3-9-2017

An Analytical System for Determining Disciplinary Vocabulary for Data-Driven Learning: An Example from Civil Engineering

Philippa Jean Otto
Portland State University

10.15760/etd.3348

An Analytical System for Determining Disciplinary Vocabulary for Data-Driven Learning: An Example from Civil Engineering

by

Philippa Jean Otto

A thesis submitted in partial fulfillment of the
requirements for the degree of

Master of Arts
in
Teaching English to Speakers of Other Languages

Thesis Committee:
Susan Conrad, Chair
Alissa Hartig
Julia MacRae

Portland State University
2017

Abstract

Data-driven learning (DDL), an inductive teaching approach in which students learn through corpus interaction, has gained recent traction as way to teach specialized vocabulary in English for Specific Purposes (ESP) classes. There is little research, however, that addresses how to choose specialized vocabulary for teaching with DDL.

This study addressed this gap in research by exploring the potential of a three-part analytical, corpus-based system for determining vocabulary to teach with DDL for a specific context of language use. This system included (1) identifying words that were significantly more frequent in a specialized expert corpus than in a corpus of general English, (2) narrowing to words that showed patterned differences in use between the specialized corpus and a student corpus, and (3) narrowing further to words with salient enough patterns of usage to teach with DDL. This three-part system was applied to the context of civil engineering in order to find vocabulary words to teach civil engineering students with low-proficiency writing skills at Portland State University.

For the first step in my analytical system, I found 201 words that occurred significantly more frequently in civil engineering practitioner writing than in the Corpus of Contemporary American English and that met requirements for frequency, distribution, and other criteria. I tested the second and third steps on 45 of these words and identified 14 words that showed evidence of needing to be taught and being well suited to DDL.

After reflecting on my process, I found that the analytical system was successful in meeting my goals for finding civil engineering vocabulary for data-driven activities. I

also made several observations that may be useful for ESP teachers who are interested in applying this methodology for their classes, the most notable of which were:

1. The system was especially useful for connecting words that are not explicitly civil engineering themed (e.g., *encountered* or *using*) to important writing functions that civil engineers perform.
2. Although it provided a systematic basis for vocabulary teaching decisions, the process was generally time-consuming and required complex judgments, which indicated that it may only be worth performing if teachers plan to regularly incorporate DDL vocabulary instruction into their course.

Acknowledgments

Thank you to everyone in my life who has shown me support and encouragement during this long process. I appreciate all of you. In particular,

- Susan—Thank you for introducing me to corpus linguistics; for your guidance, patience, and encouragement; for the time you devoted to commenting and discussing drafts; and for being excited about my research when I wasn't.
- Alissa and Julia—Thank you for your comments, ideas, and suggestions in the proposal and defense stages. I couldn't have asked for a better committee.
- Lynn, April, and Jesse—Thank you for your very helpful comments on previous drafts.
- Nathan—Thank you for your constant love and support and for always encouraging me to learn and grow.

Thank you all!

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The importance of vocabulary for language proficiency is well established in second-language education research (see, e.g., Hu & Nation, 2000; Nation, 2006; Stæhr, 2008). In fact, reports of learners' vocabulary needs are often daunting. Hu and Nation (2000) reported that adequate reading comprehension requires knowing 98% of the words in a text. Nation (2006) estimated that learners need a vocabulary of about 8,000–9,000 word families for general reading comprehension and about 6,000–7,000 word families for general speech comprehension. Stæhr (2008) reported a high correlation between vocabulary size and writing proficiency and noted the importance of "deep word knowledge" that goes beyond knowing the meaning of a word and includes knowing "word parts, collocations, synonyms and register constraints" (p. 150). It follows, then, that vocabulary development is important for language instructors to consider when preparing students for the demands of language use.

Vocabulary development becomes even more critical in specialized contexts of language use, such as medical practice, legal practice, or business practice. The word-family estimates from Nation (2006) do not take into account the high concentration of technical vocabulary—as many as 5,000 technical words in some disciplines (Coxhead, 2013)—found in specialized contexts. Additionally, words that students already know from general contexts can take on technical meanings in specific contexts, such as with the word *solution* in the context of chemistry or the word *force* in the context of physics (Mudraya, 2004).

Beyond the numbers of words required for language use, proficiency in specialized contexts requires knowing how words are used in these contexts. Language

use in specific fields or disciplines varies according to the needs and values of those who practice in those fields or disciplines—to the point where language use is specialized on many levels, including the word-level, sentence-level, and genre-level (see, e.g., Biber & Conrad, 2009). The context of use can affect, for example, the grammatical structures in which a word occurs in or the collocations in which it is found. Additionally, some words may not be appropriate in a specific register, such as the words *guarantee* or *best* in civil engineering writing (Conrad, in press). Learning how to use specialized vocabulary is an important part of learning to communicate in a specific context (Coxhead, 2013; Woodward-Kron, 2008).

Being able to communicate effectively in a specific context is part of what is termed *disciplinary literacy*, the language abilities that allow people to "create, communicate, and use knowledge" within a discipline (Shanahan & Shanahan, 2012, p. 8). The need for disciplinary literacy affects both first- and second-language university students because general language proficiency is not sufficient for the specific demands of an unfamiliar discipline. Additionally, as many have pointed out, traditional categorizations such as *first-language students* versus *second-language students* or *home students* versus *international students* often miss students who fall somewhere in the middle, such as Generation 1.5 students, or students without access to sufficient disciplinary literacy education before beginning college (Harklau, 2000;Wingate, 2015). The present study is primarily situated in English for Specific Purposes (ESP) research intended for second-language education, but the need for disciplinary literacy affects a wider population and specialized vocabulary instruction can benefit students from a range of language backgrounds.

1.1 Data-Driven Learning for Teaching Specialized Vocabulary

Given the importance of context-specific vocabulary development, it is critical for instructors to teach vocabulary in a way that exposes students to how words are used by members of specific disciplinary communities. Data-driven learning (DDL), a discovery-based approach to language teaching, is ideally suited to this goal. In DDL, students interact with language data from a corpus, often in concordance form, in order to induce rules and information about language use (Smart, 2015). A concordance list displays occurrences of a target word or phrase in the center of a line with some context on either side (see Figures 1.1, 2.1, and 2.2 for examples of concordance lines). DDL's strength for teaching specialized vocabulary use lies in the fact that interacting with corpus data exposes students to a large number of naturally occurring examples of a word in the specific context of language the students are studying. The corpus tools used in DDL present these examples in a condensed and ordered format that allows students to notice trends in how a word is used.

Figure 1.1 illustrates one way that DDL can be implemented with an excerpt from an activity used in Yunus and Awab (2014) for a Legal English class. In this activity, the instructor prepared a prompt question and included three concordance lines for students to use to answer the question. The instructor also highlighted the target word *binding* with bold typeface and highlighted recurring *be* verbs by putting them in parentheses. DDL has been used in a wide variety of ways (see Chapter 2) with the common purpose of helping students to take an active role in figuring out how language works in a specific context and to apply their findings to their own language use in that context.

4. What is the part of speech of the word ***binding*** in the lines below?
Concordances: Taken from the Law of Contract Corpus (LCC)
14. which are made and are not intended (to be) rigid, **binding** arrangements. Salmon LJ
16. he presumption that it was intended (to be) legally **binding**. The Court of Appeal
29. mediately posted an acceptance which (was) held **binding** because the delay

Figure 1.1 Activity excerpt used in a Legal English class by Yunus and Awab (2014)

1.2 The Problem of Choosing Specialized Vocabulary for DDL Activities

Despite its great potential, DDL can require more time than some deductive vocabulary teaching methods and often requires significant student training. Many have cautioned that it is important to provide adequate time, support, and guidance to students (e.g., Boulton, 2009; Smart, 2014; Vannestål & Lindquist, 2007). In order to take advantage of the great potential that DDL offers for vocabulary development, it is important to choose vocabulary words that are best suited to teaching with this method. This can be a difficult task, since few researchers have addressed how to choose specialized vocabulary to teach with DDL. Some researchers have reported their methods for choosing the specialized vocabulary that they taught with DDL (Graham and Osment, 2013; Hou, 2014; Huang, 2014; Yunus & Awab, 2012), but these studies generally focused on expert corpora alone to determine what words were important to a specific context, or—in the case of Yunus and Awab—only looked at student errors to determine what words to teach. Only one of these studies examined a learner corpus to see which words students needed to learn, and none of these studies addressed which words are learnable through DDL and which words might be better taught with a deductive approach.

The need for a way to identify specialized vocabulary for DDL activities became evident to me when I was trying to create activities to use with civil engineering students. I had some familiarity with civil engineering writing from my work as a research assistant with the Civil Engineering Writing Project (http://www.cewriting.org), but—like many ESP instructors—I did not have expertise in the content field for which I was teaching. I was unsure where to start, but I wrote a list of potential words and phrases that I had seen commonly or noticed student errors in:

- consists of, consisting of
- based on/based off
- analysis, analyses, analyze, analysis
- range from, range of, ranging from
- recommend [noun phrase], recommend that

Looking at these initial vocabulary choices, I found myself with even more questions: Would these words seem too obvious to the students in my workshops? Would the students be able to learn useful information about them through DDL? Were these words important enough to civil engineers to be worth taking the time to teach through DDL? It seemed to me that I had no real basis for choosing these words other than having happened to notice them.

I set out to find a more systematic way to identify vocabulary to teach through DDL, which led this study.

1.3 The Present Study

In order to address the gap in research and my own need to choose vocabulary words for DDL, I developed an analytical, corpus-based system to identify specialized vocabulary that are important for a particular context of language use, problematic for

students, and well suited to teaching through DDL. Using civil engineering as the disciplinary context, this study explored the potential of using a systematic approach with three steps:

1. Identifying words that are significantly more frequent in a specific context compared to general writing.
2. Comparing expert and student usage to determine which words are problematic for students.
3. Examining how salient the usage patterns are in the expert corpus to determine which words would be well suited to teaching through DDL.

This study also reviewed and assessed the effectiveness of the process in order to inform instructors who may be interested in using this tool to create data-driven vocabulary activities for their own teaching contexts and disciplines.

1.4 Overview of the Thesis

Chapter 2 begins with a discussion of the importance of disciplinary literacy, presents DDL as a strong approach for teaching specialized vocabulary, reviews previous literature, and highlights the need for my study. Chapter 3 presents the methodology that I used in testing my analytical system for determining vocabulary to teach with DDL in ESP classes and describes how I assessed the effectiveness of the system. Chapter 4 reveals the findings of this study in two parts: the first part presents the results of applying the analytical system to the context of civil engineering, and the second part discusses my findings related to the effectiveness of the analytical system. Chapter 5 concludes the thesis and offers directions for future research.

Chapter 2: Literature Review

This chapter reviews relevant research in order to establish the context and need for my study. My review of the literature begins by defining and discussing a foundational concept for this study: disciplinary literacy. I define disciplinary literacy and discuss how disciplinary literacy—and vocabulary development in particular—is important for both first- and second-language speakers of English. Next, I present data-driven learning (DDL) as a strong approach for teaching specialized vocabulary and discuss how instructors have addressed some of the challenges of implementing DDL. Finally, I discuss the gaps in how previous research has determined what specialized vocabulary to teach with DDL, and I highlight Engineering English as an area where more research would be useful. I close the chapter by introducing the present study, which used an analytical system for determining disciplinary vocabulary to teach with DDL.

2.1 The Importance of Disciplinary Literacy and Vocabulary Development

The concept of teaching literacy for specific contexts goes by a variety of names—disciplinary literacy, academic literacy, special-purposes English—and is discussed in a variety of educational contexts—from middle-school and high-school reading (e.g., Shanahan & Shanahan, 2012) to university writing (e.g., Henderson & Hurst, 2006) to second-language English for Specific Purposes (e.g., Weber, 2001). In this thesis, I use the term *disciplinary literacy* to refer somewhat broadly to the language abilities necessary to participate competently in a specific academic discipline, professional field, or other special-purpose context.

The belief underlying a focus on teaching disciplinary literacy is that language is used differently in different contexts, and that its use is shaped by the purposes and functions of the specific fields, disciplines, or discourse communities that use it (Biber & Conrad, 2009; Shanahan & Shanahan, 2012). When someone enters a new context of language use—such as when a student first starts college or begins pursuing a career in nursing—many of the tools for communicating and engaging with content are new: there is unfamiliar vocabulary and different expectations for textual organization, grammar choices, and other features of use.

Vocabulary development in particular is widely acknowledged as important for disciplinary literacy (e.g., Chujo and Utiyama, 2006; Coxhead, 2013; Shanahan & Shanahan, 2012). There are a large number of specialized vocabulary words for each discipline—perhaps as many as 5,000 technical words in some disciplines, not to mention vocabulary words that are nontechnical but take on particular meaning or usage in a discipline (Coxhead, 2013). Learning the core concepts of a discipline is tied to learning and understanding the vocabulary to describe those concepts. There are also social effects of learning disciplinary vocabulary: being able to use special-purposes vocabulary is key to learners' participation and sense of belonging within a discipline (Coxhead, 2013).

Recognizing that different contexts create different literacy needs, many educators focus on identifying and teaching—as Shanahan and Shanahan (2012) described—"the knowledge and abilities possessed by those who create, communicate, and use knowledge within the disciplines" and "the unique tools that the experts in a discipline use to engage in the work of that discipline" (p. 8). These abilities and tools are critical for students to be able to learn disciplinary content and demonstrate understanding (Woodward-Kron,

2008), and to become "fully-fledged members" of a disciplinary community (Coxhead, 2013, p. 116). Content teachers are often unaware of how to explicitly teach the language of their discipline, and language teachers do not typically have expertise in the discipline that they are preparing students for. Thus, linguistic research is important for disciplinary literacy education because it reveals the language abilities, tools, and vocabulary used in specific discourse communities (Brozo, Moorman, Meyer, and Stewart, 2013; Moje, 2008).

Because general competence in reading, writing, and speaking is not sufficient to be able to fully engage in a discipline, disciplinary literacy is an issue that affects both first- and second-language speakers. However, in American and UK university settings, academic and disciplinary literacy education and support tends to be targeted at second-language speakers alone (Wingate, 2015). Wingate argued that targeting second-language speakers is problematic because it assumes that academic literacy is only lacking in students who are traditionally categorized as second-language learners, and because it neglects other students who need literacy support. Wingate explained that as university populations grow more diverse, there is increasing need for academic literacy education. Many students come from educational backgrounds that have not adequately prepared them for the academic and disciplinary literacy demands of university courses. Students also come from a range of language backgrounds. Many students, such as Generation 1.5 students who immigrated to the U.S. when they were young, cannot be neatly categorized as first-language or second-language speakers. The need for disciplinary literacy education at the university level affects a wide group of students.

Despite the wide need for disciplinary literacy at the university level, most of the research in the remainder of this literature review is from the context of ESP. This imbalance in the research highlights the gap in literacy support for non-second-language university students and draws attention to research tools and teaching approaches from second-language teaching that have the potential to enrich the broader field of disciplinary literacy education. Because university-level disciplinary literacy education exists primarily in the context of ESP, my thesis mainly approaches teaching specialized vocabulary in the context of ESP and explores tools for teaching vocabulary that are useful to ESP teachers. In Chapter 5, I revisit the wider context of disciplinary literacy and the implications of my research for teachers outside of a second-language education context.

2.2 Data-Driven Learning for Teaching Specialized Vocabulary

One recent approach to teaching specialized vocabulary is data-driven learning (DDL). Since Johns (1991) named this approach, DDL has been growing in popularity with second-language instructors, especially for teaching ESP. Recent studies have looked at the use of DDL in ESP contexts such as English for Academic Purposes (e.g., Charles, 2014; Gaskell & Cobb, 2004; Yoon, 2008), English for Medical Purposes (e.g., Donesch-Jezo, 2010; Önder, 2014), Engineering English (e.g., Mudraya, 2004, 2006), Legal English (e.g., Weber, 2001; Yunus & Awab, 2012, 2014), and English for Tourism (e.g., Hou, 2014; Marinov, 2013; Marzá, 2014).

DDL is a discovery-based approach to language teaching in which learners interact with language data from a corpus to find patterns and form their own

generalizations or rules about usage. DDL draws from corpus linguistics, an approach to language research that involves computer-assisted analysis of large collections of purposefully chosen, naturally occurring language. The precision afforded by a corpus makes it a powerful tool for ESP. Corpora can be designed to focus on specific language varieties so that the resulting concordance data reflects language use in a specific context or domain. Corpus linguistics research has made many contributions to disciplinary literacy education (e.g., Conrad, Pfeiffer, & Szymoniak, 2012; Kaewpet, 2009; Salager-Meyer, 1994), but DDL brings the corpus research process into the classroom by teaching learners to investigate language for themselves (e.g., Gaskell & Cobb, 2004; Liu & Jiang, 2009; Sun & Wang, 2003; Yoon, 2008). Students can use corpora to discover patterns of language use and natural language data that are directly applicable to their fields of study or career goals.

2.2.1 DDL's strengths for teaching specialized vocabulary. DDL has a number of strengths that align with current approaches to language teaching, including integration of lexicon and grammar, consciousness-raising, inductive learning, and teaching grammar in discourse contexts. These strengths make it a powerful tool for teaching students how to use specialized vocabulary effectively. There is considerable overlap between these concepts, but it is worthwhile to consider the relationship of DDL to each concept individually.

One concept that is key to understanding DDL's strength for teaching vocabulary is the integration of lexicon and grammar. Corpus research has played a role in demonstrating how the vocabulary words and grammar are complexly interconnected; knowing a word is inseparable from knowing how it behaves grammatically in a sentence

(see, e.g., Conrad, 2000). A lexico-grammatical approach teaches vocabulary and grammar together.

DDL allows teachers to simultaneously teach vocabulary words and how these words are used in context. Liu and Jiang (2009) described the connection between lexico-grammatical perspective and DDL as inherent and suggested that concordancing can increase students' lexico-grammatical competence. An example with the word *base* illustrates how concordancing can reveal lexico-grammar. The verb *base* is an important word in both academic English and English for science and engineering. In order to use it competently, it is necessary to know information about what prepositions typically follow this verb, how to use it in the passive voice, and so on. Figure 2.1 shows concordance lines from civil engineering for the word *based*, generated by AntConc (Anthony, 2015). From this short concordance sample, a teacher or learner might notice that the prepositions *on* and *upon* seem to commonly follow the word *based* and that *based* is often preceded by a *be*-verb, making it passive. These patterns could be confirmed by further investigation with the corpus.

Another important concept in language teaching is consciousness-raising. Schmidt (1990) emphasized the importance of consciousness, or learner attention, for second language acquisition. Schmidt linked noticing (when a person pays attention to the stimulus they are perceiving) to the concept of intake, suggesting that noticing is critical for allowing learners to truly benefit from language that they are exposed to. Ellis (2002) called for "discovery-type grammar tasks" that promote noticing and raise students' consciousness of the grammatical structure of language (p. 176).

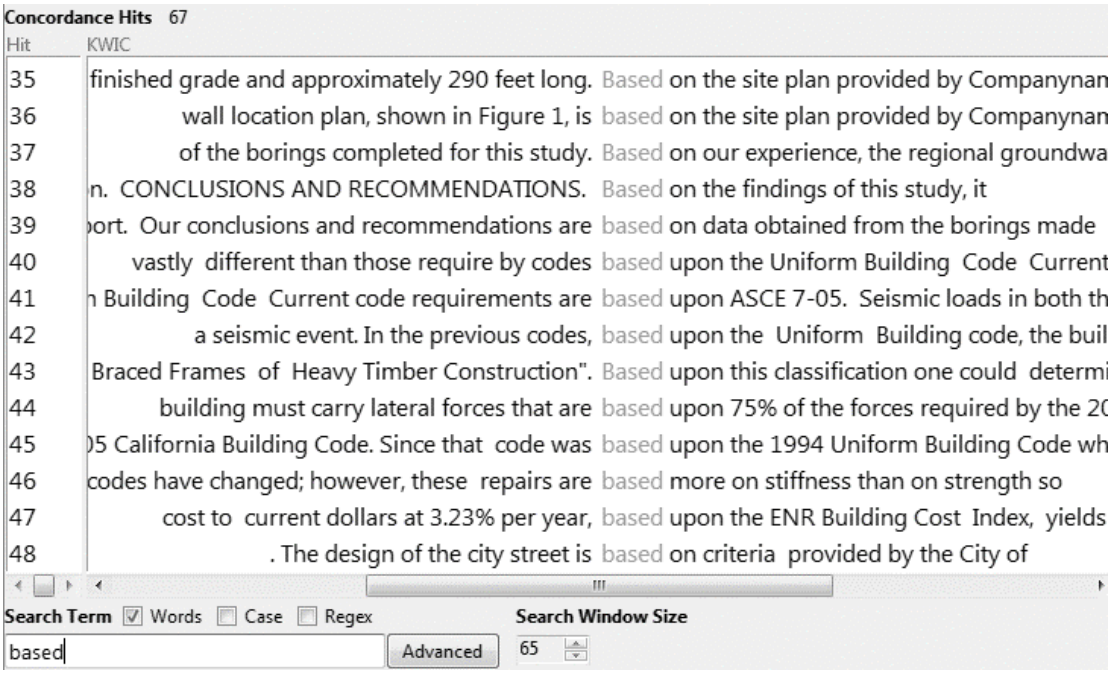


Figure 2.1 Key-word-in-context concordance lines for the search term *based*

Discovery and noticing are a core part of DDL. Concordancing software is designed to support noticing by distilling a large amount of data into a list that can be easily scanned visually for lexical patterns and by revealing which word combinations are most frequent. A concordance positions a lexical item as the visual center and draws attention to words that repeatedly occur before or after the target word. Figure 2.2 shows a concordance list for the phrasal verb *get away*, using the concordancer MonoConc Pro (Barlow, 2002) and the Louvain Corpus of Native English Essays (Centre for English Corpus Linguistics, n.d.). In MonoConc, the concordance list displays the search terms (*get away)* in blue and frequently co-occurring words (*they can*, *with it*, etc.) in red, making it easier to spot patterns. Once learners identify a pattern, they can examine the context to discover the difference between phrases like *get away with* and *get away from*.

Concordancing can raise students' awareness to aspects of the language that would take much longer to notice in unenhanced input.

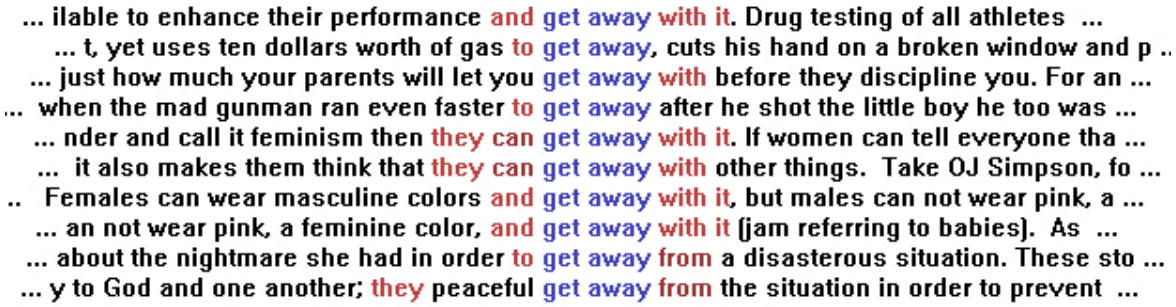


Figure 2.2 Concordance data highlighting collocations of *get away,* using MonoConc Pro (Barlow, 2002)

Inductive learning is another concept that has been gaining popularity over the last 20 years. Gollin (1998), for example, suggested that learners may more fully understand and remember the language rules they discovered themselves than rules provided by an instructor. In DDL, learners apply inductive reasoning: they are given data (examples of a phenomenon) and must work backwards to formulate a rule that governs the data. Once students discover a rule, they can apply it by producing language using their newly-formulated rule. In contrast, deductive learning begins with the instructor providing a rule (usually accompanied by supporting examples) and then asking students to apply the rule. Inductive learning can lead to greater motivation and retention of material because of students' active role in the learning process and because they have to engage more deeply with the material to induce rules (Liu & Jiang, 2009).

Sun and Wang (2003) explored this inductive/deductive distinction by contrasting the use of an inductive DDL activity to a comparable deductive activity. Both activities were informed by the same corpus data, but the inductive DDL group used a

concordancer to analyze the data, induce a rule, and apply it, while the deductive group was given a rule first with supporting example sentences (from the corpus data) and asked to apply the rule. They found that the DDL inductive group performed significantly better than the deductive group.

Additionally, teaching language in a meaningful context has been a central issue in current language teaching approaches, such as communicative language teaching (see Brandl, 2008). Particularly in grammar teaching, there has been increased interest in explicit "focus on form" instruction that emphasizes raising students' awareness of grammatical forms as they occur in their discourse context—that is, preserving the meaning and function of the complete utterance (Brandl, 2008; Conrad, 2000; Ellis, Basturkmen, & Loewen, 2002). Focus on form teaches form, meaning and use simultaneously by calling attention to a specific form as a necessary tool for successfully completing a communicative task (Ellis, Basturkmen, & Loewen, 2002).

DDL aligns with a focus on form approach because it processes a large amount of input and calls attention to specific grammatical forms that are occurring in language that was produced for communicative purposes. Corpora are typically built using data in the form of full discourses or communicative events—essays, letters, work memos, transcripts of full conversations, and so on. A concordancer takes these data and organizes the language into KWIC lists that show the immediate context of the word. Once viewers identify a pattern or sample to investigate further, they can view the fuller context and identify the communicative functions that the structures accomplish.

2.2.2 Addressing DDL's challenges. Despite the strengths that DDL can offer to the classroom—integration of lexicon and grammar, consciousness-raising, inductive learning, and teaching grammar in discourse contexts—a significant criticism and recurring caution is that corpus data can be overwhelming or confusing to students (see, e.g., Smart, 2014; Vannestål & Lindquist, 2007). With a large corpus, student concordancing can uncover too many occurrences of a search term to process or make sense of. With a smaller corpus, students may not find enough occurrences of their search term to make a generalization or may make an inaccurate generalization. Additionally, some lexico-grammatical patterns are more difficult to induce than others (see, e.g., Sun and Wang, 2003). Instructors have addressed the challenges that the DDL poses for students by implementing DDL in different forms based on the needs of their students.

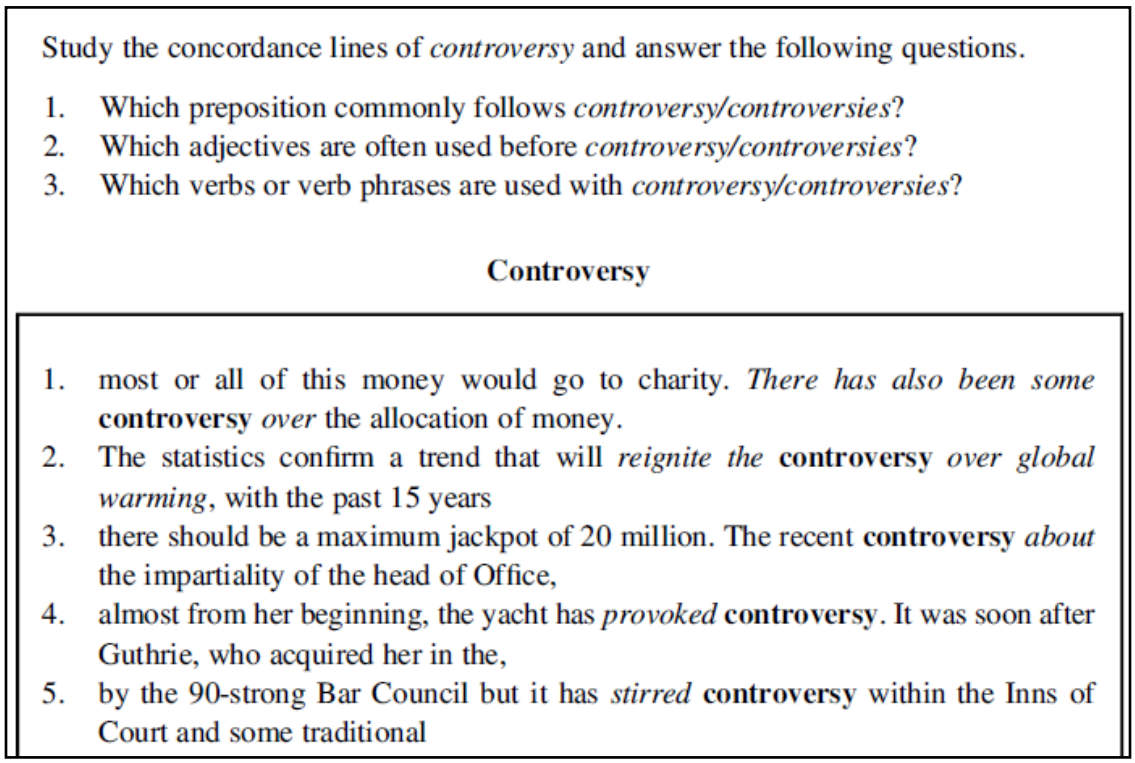
One way that DDL can be modified based on student needs is with the form of contact students have with corpus data: DDL can be either "hands-on" (also called "soft" DDL; see, e.g., Charles, 2014; Yoon, 2008) or "paper-based" (also called "hard" DDL; see, e.g., Huang, 2014; Smart, 2014). Hands-on DDL typically involves students using a concordancer to investigate a language question. A concordancer is a software tool that makes the data in a corpus useful by organizing and presenting them accessibly. It generates a list of all occurrences of a search term in their immediate context (see Figures 2.1 and 2.2 for example concordance lines). Searchable online corpora such as the Corpus of Contemporary American English (Davies, 2008–) have built in concordancing features, but concordancing software such as MonoConc Pro (Barlow, 2002) or AntConc (Anthony, 2015) can also be used to concordance any chosen or self-compiled corpus. Other corpus tools or techniques frequently used in hands-on DDL include word

frequency comparisons (see Figure 2.3, for example), cluster or n-gram tools (tools that identify formulaic sequences), and collocate tools (tools that identify the strongest or most frequent collocates for a given word).

WORD 1 (W1): BIG (1.68)			WORD 2 (W2): LARGE (0.60)		
	WORD	SCORE		WORD	SCORE
1	BROTHER	1,928.2	1	SAUCEPAN	2,320.7
2	LEAGUES	1,011.2	2	NONSTICK	1,494.5
3	TROUBLE	969.5	3	EXTENT	649.0
4	SURPRISE	802.7	4	HADRON	550.8
5	NAMES	706.3	5	MAGELLANIC	413.1
6	SISTER	699.1	6	QUANTITIES	359.8
7	DEAL	538.7	7	SKILLET	335.9
8	LEAGUE	488.3	8	INTESTINE	285.5
9	SUR	481.2	9	YOLKS	278.8
10	HUG	444.2	10	PROPORTION	267.0
11	MAC	426.4	11	SAMPLES	258.6
12	WINNER	356.1	12	SAUTÉ	241.8
13	DIPPER	334.7	13	BAKING	224.2
14	DADDY	289.4	14	CARROTS	218.3
15	SECRET	289.4	15	MIXING	212.4

Figure 2.3 Comparison of collocates for *big* versus *large*, ordered by ratio score, using COCA's online Compare tool (Davies, 2008–)

Paper-based DDL involves giving students copies of already compiled and printed concordance data to work with (Huang, 2014). Teachers choose data that will help students see a pattern and typically create a handout or worksheet to guide students through the activity. The example in Figure 2.4 shows an activity from Huang's study that focuses on the word *controversy.* In this example, the instructor has provided concordance lines (inside the box) and questions to guide students' attention (above the box). The instructor has also added italicization to highlight collocations or sequences that the students should notice.



Study the concordance lines of *controversy* and answer the following questions.

1. Which preposition commonly follows *controversy/controversies*?
2. Which adjectives are often used before *controversy/controversies*?
3. Which verbs or verb phrases are used with *controversy/controversies*?

Controversy

1. most or all of this money would go to charity. *There has also been some* **controversy** *over* the allocation of money.
2. The statistics confirm a trend that will *reignite the* **controversy** *over global warming*, with the past 15 years
3. there should be a maximum jackpot of 20 million. The recent **controversy** *about* the impartiality of the head of Office,
4. almost from her beginning, the yacht has *provoked* **controversy**. It was soon after Guthrie, who acquired her in the,
5. by the 90-strong Bar Council but it has *stirred* **controversy** within the Inns of Court and some traditional

Figure 2.4 A paper-based data-driven activity used in Huang's (2014) study

There is ongoing discussion of whether hands-on applications of DDL are effective—and for what levels of language proficiency these approaches are appropriate. Boulton (2009) argued that the training required for student concordancing can be a barrier to learning, especially for low-level learners, since the inductive process itself takes a significant amount of time for students to get used to. Smart (2014) echoes this assessment, saying that when learners are unfamiliar with both the software and the discovery-based approach, hands-on concordancing can overwhelm and detract from the learning experience. Boulton and Smart both advocated a paper-based approach.

Vannestål and Lindquist (2007) also acknowledged the amount of time required for students to get used to working with a concordance, but their solution was to start with paper-based materials to get students used to the inductive process of working with

corpus data, followed by significant training time for students to learn how to use the concordancing software. Other studies, such as Lee and Swales (2006), Yoon (2008), and Charles (2014), have reported successful implementations of hands-on DDL, often (but not exclusively) with more advanced (both in language proficiency and in education level) students.

Another way that the challenges of DDL can be addressed is through scaffolding. The level of scaffolding, or instructor guidance, in DDL can vary. In some cases, instructors have a defined linguistic goal: they define a question (that they already know the answer to) and lead students to discover this answer using a structured handout (e.g., Huang, 2014). In other cases, the instructors let students choose their own questions to investigate, not knowing ahead of time what their students will discover (e.g., Yoon, 2008). Smart (2014) advocated for the practice of "guided induction." He differentiated between a "purely inductive" approach and that of guided induction, describing guided induction as a scaffolded framework in which students can make discoveries (pp. 186–187).

Regardless of the form of DDL—hands-on or paper-based—or the level of instructor guidance, DDL generally takes time and commitment to train students to extract information that will be useful for their language development from the corpus data. Even with training, students often find DDL a time-consuming way to learn (Kennedy & Miceli, 2001). Given the challenges involved in DDL and the time investment that it represents for both instructors and students, it is important that DDL be implemented thoughtfully and intentionally.

2.3 The Problem of Deciding What to Cover in a DDL Lesson

An important part of implementing DDL thoughtfully and intentionally is choosing the most effective content to teach. Although many studies have reported using DDL to teach the specific lexical items (e.g., Graham and Osment, 2013; Hou, 2014; Huang, 2014; Marzá, 2014; Önder, 2014; Yunus & Awab, 2012), fewer have addressed how they chose the words they decided to teach (Graham and Osment, 2013; Hou, 2014; Huang, 2014; Yunus & Awab, 2012). Of the four studies that reported how they chose vocabulary, all but Yunus and Awab (2012) relied on a measure of frequency to identify words that were important to the context of use.

Graham and Osment (2013) created activities for engineering students using a corpus of engineering textbooks. Graham and Osment chose words for these activities by generating a raw wordlist from the engineering textbook corpus ordered by frequency, removing non-content words and words that occurred in fewer than 10 textbooks, organizing the list into word families, and then intuitively choosing 12 words that were highly frequent and had particular usage in engineering contexts.

Hou (2014) created DDL activities for an English for Hospitality class on winetasting. Hou built two corpora of wine tasting notes from the Liquor Control Board of Ontario: one for red wine and one for white wine. Hou created lists of specialized vocabulary by generating a raw wordlist, deleting function words and words that belonged to the General Service List (GSL) or Academic Word List (AWL), and picking the 20 most frequent of the remaining words from each corpus. Hou confirmed the usefulness of these words by showing the lists to wine-tasting experts. Hou also reported performing keyword analysis to determine which words occurred significantly more

frequently in each of the wine corpora than in the Spoken British National Corpus, but it was not clear how the keyword analysis factored into the decision about which words to teach.

Huang (2014) focused on teaching how to use specific abstract nouns in academic opinion essays. Huang built a topic-specific corpus of texts related to the lottery (a topic students were preparing to write about) and used keyword analysis to identify words that were significantly more frequent in the specialized corpus compared to the British National Corpus. From this keyword list, Huang chose abstract nouns commonly used in opinion essays that occurred at least three times in the lottery corpus.

Graham and Osment (2013), Hou (2014), and Huang (2014) all used corpora designed to represent the target language variety; however, none of them looked at learner language. Learner corpora are an important part of assessing needs for ESP. While expert corpora allow researchers to determine the characteristics of a target register, it does not always follow that learners need to learn all of these characteristics equally, or that all of the vocabulary identified in these studies are equally problematic for students. Gilquin, Granger, and Paquot (2007) emphasized the importance of learner corpora in English for Academic Purposes for measuring the level of difficulty that learners have with different aspects of language. They explained that the majority of corpus research for English for Academic Purposes has focused on native or expert corpora alone, and that without learner corpora, teachers and material designers tend to rely on intuition for deciding what to teach.

Yunus and Awab (2012)—unlike Graham and Osment (2013), Hou (2014), and Huang (2014)—did address specific learner needs when choosing vocabulary to teach

through DDL. They chose collocations to teach with DDL in Legal English classes based on identifying the most frequent collocation errors in a set of essays students wrote previous to the study. Yunus and Awab did not address, however, how to tell which of the collocations were important for Legal English specifically and whether they relied on their experience with Legal English in choosing the collocations to teach.

Additionally, none of the four studies (Graham & Osment, 2013; Hou, 2014; Huang, 2014; Yunus & Awab, 2012) addressed how to determine whether or not words would work well in DDL activities. Some studies have commented on the level of difficulty of certain structures and the effect on whether students were able to learn them. Sun and Wang (2003), for example, found that after DDL activities, students only made significant improvements on error correction tasks with easier collocation patterns, suggesting that easier patterns might be more compatible with an inductive approach. However, Sun and Wang identified these easy and difficult patterns by consulting experienced instructors and did not base these categorizations on whether the words were easy or difficult to learn specifically with DDL. Smart (2014) similarly commented on not all structures being equally well suited to DDL, but did not describe what makes structures well suited to teaching through DDL. While there is relatively little research on what exactly makes a structure more or less suited to teaching through DDL, it makes sense that since time constraints limit how much can be taught with this approach, teachers should choose structures carefully to ensure that students are neither overwhelmed by trying to find patterns for words that occur in overly complex structures nor feel like their time is wasted by investigating overly simple vocabulary patterns that could have been more efficiently explained deductively.

There is a need for more research that addresses how to identify ESP vocabulary to teach with DDL, taking into account learners' individual vocabulary needs and vocabulary words' suitability to DDL.

2.4 Engineering English and DDL Vocabulary Development

One area of ESP where more research on DDL and specialized vocabulary instruction would be usefully applied is Engineering English. A number of recent researchers have drawn attention to communication skills, and particularly writing, as vital to engineering practice (Nelson, 2000; Swarts & Odell, 2001; Winsor, 1990; Yalvac Smith, Troy, & Hirsch, 2007). Swarts and Odell (2001) pointed out that engineers are dependent on each other's work and that effective communication is essential to engineering practice. Winsor (1990) noted that engineers often take writing for granted but that, in fact, writing is a significant tool that mediates engineers' knowledge (i.e., their knowledge does not come from lab work, but from the documentation of lab work) and allows communication of this knowledge.

Not only is communication important for engineers, but the field of engineering also has register- and genre-specific conventions that need to be learned in order to succeed in engineering practice. Nelson (2000) described this need succinctly: "Successful engineering writing entails adapting to a new discourse community and mastering its conventions for communication" (p. S2B-2). Knowing general English is not sufficient for successful communication in the engineering profession.

Engineering English is also an area where DDL is promising, not only because of DDL's strength in tailoring to a specific register, but also because of the nature of engineering students themselves. Mudraya (2004, 2006) argued for using DDL with

engineering students for both these reasons: she pointed out that the engineering field has specific language requirements, such as specialized vocabulary, and that engineering students' analytical and technological skills make them well suited to the data-driven approach.

Mudraya (2006) explored vocabulary usage in engineering textbooks and noted the importance of both technical and nontechnical vocabulary to engineering. She argued that nontechnical or "subtechnical" vocabulary—words like *solution* that have both common and technical uses—require more work to learn than technical vocabulary like *vulcanize* that have more clearly defined usage and can be easily glossed. Concordancing activities with a specialized engineering corpus can help students notice how vocabulary words—even words they already know—are used specifically in the engineering field.

Mudraya (2004) argued that DDL is especially well fitted for engineering students because their strengths include analytical thinking and technical expertise. Common roadblocks to student concordancing are the learning curve for using the software and the unfamiliar process of inducing patterns from data (Boulton, 2012; Vannestål and Lindquist, 2007; Yunus & Awab, 2014), but Mudraya argued that the technical aptitude of engineering students would be well suited to hands-on DDL activities. Mudraya also argued that a data-driven approach would be more enjoyable for analytically-minded engineering students than a traditional, teacher-fronted approach.

2.5 The Present Study

As the above sections have argued, teaching disciplinary literacy—and particularly specialized vocabulary—is important for students navigating a new field of study and for preparing to communicate in a specific occupational field. DDL is a strong

tool for teaching specialized vocabulary because it can reveal word usage in specific contexts and because of its grounding in language teaching principles. Despite its great potential, DDL can require extra time from instructors and students, so it is important to choose the most effective content to teach with this approach. While a few DDL studies have described their methods for determining what vocabulary to teach in particular ESP contexts, there is a need for more studies that examine how to choose vocabulary for ESP-focused DDL materials, and, in particular, that examine both expert and learner corpora and apply pedagogical judgment to address whether a word is well suited to DDL. The area of Engineering English is especially ripe as a context for this exploratory research.

To address the question of how to choose vocabulary, I created an analytical system for determining vocabulary to teach with DDL in an ESP course. The purpose of this thesis was to investigate the potential for using this systematic process to:

- Identify words that were significantly more frequent in the ESP content area than in general writing that students are likely to be familiar with.
- Narrow to words that are likely to be problematic for low-proficiency students, who are likely to benefit from language support.
- Narrow further to words likely to be well suited to teaching through DDL activities rather than being taught deductively.

More specifically, I used civil engineering as the disciplinary context (for reasons described in Chapter 3) to investigate the combination of corpus linguistics tools and pedagogical judgment that met the goals for the system. Then, following the

investigation of civil engineering vocabulary, I reviewed the process and the benefits and challenges it would present if replicated by other teachers or in other fields.

Chapter 3: Methodology

This chapter presents the methodology for this study that explored the usefulness of an analytical system for determining vocabulary to teach with data-driven learning (DDL) in an English for Specific Purposes (ESP) course. The following sections describe the context that I chose for applying this system, the corpora that I used, the process of applying the system, and how I assessed the effectiveness of the system.

3.1 The Context of the Study

The analytical system that I developed was designed to customize vocabulary choices for a specific area of ESP and a specific group of learners. To test the system, I chose the context of Civil Engineering English and I chose to target the needs of undergraduate students with low-proficiency writing skills in the Civil and Environmental Engineering program at Portland State University (PSU). These contexts were chosen for this study mainly because of my involvement as a research assistant in the Civil Engineering Writing Project (a larger research project being conducted at PSU) and the availability of previously-collected civil engineering texts from this larger research project. In the following sections, I describe in more detail two important areas of context for this study: the Civil Engineering Writing Project and PSU's Civil and Environmental Engineering program.

3.1.1 The Civil Engineering Writing Project. The Civil Engineering Writing Project (http://www.cewriting.org) has been investigating the gaps between writing in civil engineering practice and the writing students produce for class assignments. They have been creating and piloting materials to address these gaps (see, e.g., Conrad, 2015;

Conrad & Pfeiffer, 2011; Conrad, Pfeiffer, & Szymoniak, 2012). Since 2009, the Civil Engineering Writing Project has collected over 500 practitioner documents and over 1000 student papers for its still-growing corpus. The present study used a subset of these texts for analysis.

In addition to providing texts, the Civil Engineering Project is also significant as context for the present study because the project has explored and documented the importance of writing for civil engineering and the need for student writing instruction. On a basic level, writing is important in civil engineering because conveying information inaccurately or ambiguously can lead to serious financial loss, injuries, or deaths (Conrad, Pfeiffer, & Szymoniak, 2012). More than that, however, the role of writing is deeply integrated with the practice of civil engineering; by combining textual analysis with interviews of practitioners and students, the Civil Engineering Project has highlighted this integration. Conrad (in press) observed consistent patterns in practitioner writing in terms of word choice, sentence structure, and textual organization—and found that student writing differed significantly in these areas. Conrad (to appear) showed how differences in practitioner and student writing, specifically with passive voice, were linked to students' lack of understanding of key values in civil engineering for clear, precise, and accurate content; concise, easy reading for clients; and liability management. These studies add evidence that writing is critical to the field and that it is not simply a skill students can acquire in general-purpose writing classes; writing is inextricably tied to the work and core values of civil engineers. With this evidence in mind, I continued to examine practitioner and student writing in this study, focusing specifically on

vocabulary words that are important for practitioners' work and examining how these words are used in civil engineering writing.

3.1.2 The educational context. This study focused on the Civil and Environmental Engineering undergraduate degree program at PSU, one of the programs in the Civil Engineering Writing Project. The Civil and Environmental Engineering department offers two Bachelor of Science options—civil engineering and environmental engineering—with about 430 combined undergraduates enrolled in 2015. There is considerable overlap in the courses required for the two majors, and for the purposes of this study I will not be distinguishing between the majors. Most students in the Civil and Environmental Engineering program plan to go into industry practice after graduating (Conrad, 2014, 2015).

The students in the Civil and Environmental Engineering program at PSU come from a diverse range of backgrounds: 23% are international students, 25% are under-represented minorities, 29% are women, and almost 45% are first-generation college students (Portland State University, 2017). An estimated quarter of students enrolled in the program speak English as a second language (A. Lewis, personal communication, May 19, 2016).

The writing background and proficiency level of the students in the Civil and Environmental Engineering program also varies. Some students choose to take a technical writing class to satisfy their general university writing requirements, but there are no further writing course requirements for non-transfer civil engineering students (Portland State University, 2017). To narrow the scope of the students I was targeting in

this study, I focused on students in the program who could be identified as exhibiting low-proficiency writing skills (see section 3.2.3 for more details).

3.2 The Corpora

This study used three corpora: the Corpus of Contemporary American English (COCA, Davies, 2008–), a civil engineering practitioner corpus, and a civil engineering student corpus. The corpora in this study were intended to represent three contexts of writing: (1) general writing from a range of contexts, which provides a basis for establishing what words are important for civil engineering writing in particular; (2) professional civil engineering writing, which sets a standard or goal for vocabulary instruction; and (3) low-proficiency undergraduate civil engineering writing, which establishes the current needs of the students that vocabulary instruction is intended to benefit. The following sections describe these three corpora in more detail.

3.2.1 The Corpus of Contemporary American English. I used COCA (Davies, 2008–) to provide broad representation of general English writing for comparison with the more specific context of civil engineering writing. Its purpose was to provide a point of reference to establish which of the words in the specialized, civil engineering practitioner corpus were significantly more frequent than in general writing.

COCA is currently the largest freely-available English-language corpus and includes written and spoken texts from a variety of genres and academic disciplines (Davies, 2008–). I used a downloaded version that included texts from 1990–2012, and I used only the written portions of the corpus in order to make a stronger comparison with the practitioner and student writing corpora. The written COCA corpus included texts

from four categories of writing: Academic, Fiction, Newspaper, and Magazine (Table 3.1). The total COCA word count used in my analysis was just over 335 million words.

Table 3.1 Summary of the Corpus of Contemporary American English written portions, 1990–2012 (Davies, 2008–)

Register	**Number of Words**
Academic	82,287,233
Fiction	77,061,285
Newspaper	88,021,009
Magazine	88,533,539
Total:	**335,903,066**

3.2.2 The Practitioner Corpus. The purpose of the practitioner corpus was to represent the typical writing of experienced civil engineers. In order for the vocabulary to be useful for students going into various areas of civil engineering and preparing for various kinds of writing tasks, the corpus needed to include writing from a range of contexts (registers) and from a range of specialties within civil engineering. Additionally, because this corpus was used to systematically compare vocabulary frequencies with a general English corpus, it was important that it be designed to balance the various registers and specialties as much as possible.

The practitioner corpus was composed of texts that had been written by licensed practicing engineers with at least five years of experience, had undergone peer review, and had been sent to real clients. This corpus included texts from 36 consulting firms in California (15 firms), Oregon (13 firms), Michigan (6 firms), Idaho (1 firm), and Ontario, Canada (1 firm).

The practitioner corpus included three common writing tasks from civil engineering practice: reports, technical memos, and site visit observation memos. Reports are a dominant register in civil engineering and composed the majority of the corpus. To make sure that a variety of specialties within civil engineering was represented, I included reports from five specialty areas: geotechnical engineering, transportation engineering, structural engineering, water resources engineering, and environmental engineering (for more information on common civil engineering specialties, see Civil Engineering Degree, 2017). The technical memos and site visit observation memos were mixed specialty. I included roughly 20,000 words from each of the five report registers and from technical memos and site visit observation memos for a total word count of 149,007 for the practitioner corpus (Table 3.2). Practitioner site visit observations only had 15,466 words available in the Civil Engineering Writing Project corpus, and so I included all the texts that were available.

Table 3.2 Summary of practitioner and student corpora

The Practitioner Corpus			**The Student Corpus**		
Register	**Number of Words**	**Number of Texts**	**Register**	**Number of Words**	**Number of Texts**
Reports (from 5 specialties)	110,612	44	Reports (400-level)	10,344	8
-	-	-	Lab Reports (300-level)	10,166	12
Technical Memos	22,929	17	Technical Memos (300- and 400-level)	9,623	15
Site Visit Observation Memos	15,466	25	Site Visit Observation Memos (100-level)	10,104	18
Totals	**149,007**	**86**	**Totals**	**40,237**	**53**

Despite efforts to make the practitioner corpus as balanced as possible, geotechnical engineering was most strongly represented. Table 3.3 shows the breakdown of word count and number of texts by specialty. The tech memos and site visit reports included mixed specialties but were mostly from geotechnical engineering. Including comparable numbers of words from each report specialty helped counteract the lack of specialty balance in the other two registers, but as a whole, the practitioner corpus was weighted towards geotechnical engineering.

Table 3.3 Practitioner corpus word count and texts by specialty

Specialty	**Number of Words**	**Number of Texts**
Geotechnical Engineering	48,203	42
Transportation Engineering	28,223	10
Structural Engineering	27,801	13
Water Resources Engineering	22,437	13
Environmental Engineering	22,343	8
Totals	**149,007**	**86**

3.2.3 The Student Corpus. The purpose of this corpus was to provide a way of assessing the needs of a particular group of students: PSU undergraduate civil engineering students with low-proficiency writing skills. To avoid teaching words that these students were already familiar and competent with, I wanted to be able to compare usage of specific vocabulary words between the practitioner and student corpora to determine which words the students had trouble with. My goals for this corpus were for it to represent students who would benefit from writing instruction, for it to be as comparable as possible to the practitioner corpus, and for it to cover a number of writing assignments that students typically perform.

In order to target students who would benefit from writing instruction—civil engineering students with low-proficiency writing skills—I needed to select texts that fit this category. In an ESP classroom context, ideally, a teacher would build a corpus of papers written by their own ESP students or by students in the same context as their own students. However, the previously-collected papers from the Civil Engineering Writing Project were written by undergraduate students with a wide range of proficiency levels and language backgrounds—not all of which fit the students I hoped to target. Since these student texts were not divided by language background or proficiency level, I chose to use counts of errors in standard written English as a measure of proficiency level.

For the student corpus, I selected only texts that had a high number of marked grammar errors. These errors were previously coded by researchers for the Civil Engineering Writing Project and included seven categories (defined in Table 3.4, from Conrad, Pfeiffer, & Szymoniak, 2012). I included only texts with 13 or more errors per 1000 words. To choose this cutoff, I compared the error counts of the coded files from each register. My goal was to keep the error cutoff as high as possible without excluding too many words from the corpus. The range of error counts per 1000 words varied among the student registers, with a low of zero for each register and a high of 31.11 for general reports, 89.45 for technical memos, 63.89 for lab reports, and 74.93 for site visit memos. Choosing to include texts with 13 or more errors per 1000 words allowed me to include about 10,000 words from each of the student registers. For technical memos, I included all of the texts with 13 or more errors per 1000 words; for the other three registers, I excluded between one and six texts—looking only at their word counts—in order to keep the registers to roughly 10,000 words.

Table 3.4 Error categories from Conrad, Pfeiffer, and Szymoniak (2012)

Verb Errors	Errors in tense or aspect, incorrect formation of infinitives or other verb structures (other than subject-verb agreement).
Sentence Structure	Any errors in the construction of sentences, including the structure or placement of relative clauses and "dangling modifiers."
Articles, Prepositions and other errors typical of ESL learners	Articles, prepositions, plurals, subject-verb agreement and pronoun-antecedent agreement. Although these errors are sometimes made by native speakers of English, they are characteristic of English as a Second Language learners.
Spelling and Typos	Errors related to spelling or typing that do not fall into other categories
Punctuation	Comma errors, sentence final punctuation errors, and other punctuation errors

I counted errors from the first three error categories only: verb errors, sentence-structure errors, and errors with articles, prepositions, and other features that are typically difficult for second-language speakers of English. I did not count spelling errors, typographical errors, or punctuation errors because, while these kinds of errors often detract from a writers' credibility and give readers an impression of writer incompetence, they are not necessarily markers of low-proficiency writing skills. I judged the first three categories of errors as better indicators of whether a student was proficient in using standard written English. The following examples (from texts included in my student corpus) show the kind of errors that were counted to determine whether to include a text (errors marked with italics, except for the error in sentence c, which affected the whole sentence):

a. The meeting with the client was informative, we *get* to know what the client *have* in mind. (technical memo)

b. *Compare between* all three diagrams, can give us a clearer image of the test. (lab report)

c. Issues to do with the shape that is L-shaped appearance and also the respective location of the site is said to be [address]. (technical memo)

This method of choosing texts by error count simply measured students' ability to use standard English grammar—which is only one small part of writing proficiency. However, counting errors allowed me to choose texts based on quantifiable evidence of the need for language instruction so that I could use the texts as the basis for vocabulary instruction decisions.

The second goal for this corpus—that it be as comparable as possible with the practitioner corpus—was achieved through including similar registers: reports, technical memos, and site visit observation memos. Since undergraduate students are generally not advanced enough to have specialties, I did not break down any of the registers by specialty as in the practitioner corpus. I also included lab reports as a register in the student corpus because lab reports are a highly common writing assignment and because students at PSU do not write reports until their fourth year. Civil engineers do not write lab reports in practice, but lab reports are a good example of the kind of reporting that students tend to be most familiar with. I included roughly 10,000 words from each student register for a total word count of 40,237 for the student corpus (about a quarter of the size of the practitioner corpus, Table 3.2).

3.3 The Analytical System

I developed the analytical system for determining vocabulary to teach with DDL by combining elements that were present in previous research and adding original elements that I devised to meet my goals. The first two steps in my process relied on commonly-used corpus techniques: specialized-general corpus comparison using keyword analysis (see, e.g., Chujo & Utiyama, 2006; Mudraya, 2006; Tangpijaikul, 2014) and expert-novice corpus comparison (see, e.g., Conrad & Pfeiffer, 2011; Flowerdew 2003; Hartig & Lu, 2014). There was less precedent for the third step in my system—determining whether or not the words were well suited to teaching through DDL. This step required more reflection to discern what qualities make a word learnable through DDL. I mapped out the three main steps of my system before beginning this study, but I also refined and further developed the system during the course of the study. As I became more familiar with the process, I modified the system to add structure to the analysis, to make the system more time-effective, and to make the judgments as straightforward as possible.

The system to determine vocabulary to teach with DDL to civil engineering students at PSU was performed in three main steps: (1) identifying words that were significantly more frequent in the practitioner civil engineering corpus than in general writing, (2) determining which of these words were problematic for the students, and (3) determining which of these problematic words would likely be well suited to teaching through DDL activities rather than being taught deductively. I performed my corpus analysis using AntConc, a corpus analysis software program developed by Anthony (2015) that allows users to analyze concordance lines, identify the strongest collocates

and the most frequent lexical bundles, and generate keyword lists. The following sections describe each of the three steps of corpus analysis (see Figure 3.1 for a summary).

3.3.1 Identifying words particular to civil engineering. To identify a list of words that were particular to civil engineering writing, I performed keyword analysis with AntConc to find words that were significantly more frequent in the practitioner corpus than in COCA.

Keyword analysis is useful in ESP research for identifying specialized vocabulary because it can reveal which words are especially frequent in a particular language context when compared to general English. For example, the word *the* is often not a keyword since it tends to be common in most contexts; however, *bake* would be likely to be a keyword in a corpus of recipes when compared to a corpus of general English.

Keyword analysis uses a statistical measure, log-likelihood ratio in this case, to compare the frequency of a word between a target corpus and a reference corpus. The result of the log-likelihood ratio test is a keyness score that represents the likelihood of the word occurring in the target corpus versus the reference corpus. Keyness scores of 3.84 or higher indicate a significance level of less than 0.05 and scores of 6.63 or higher indicate significance level of less than 0.01 (Anthony, 2016).

To perform keyword analysis in AntConc, I loaded the practitioner corpus as the main corpus and created a word list to establish target corpus frequencies. I chose log-likelihood, which is recommended for this software (Anthony, 2016), as the keyword generation method. I also chose to use a list of COCA's word frequencies as the reference corpus because of the processing time that would have been required to use COCA's raw

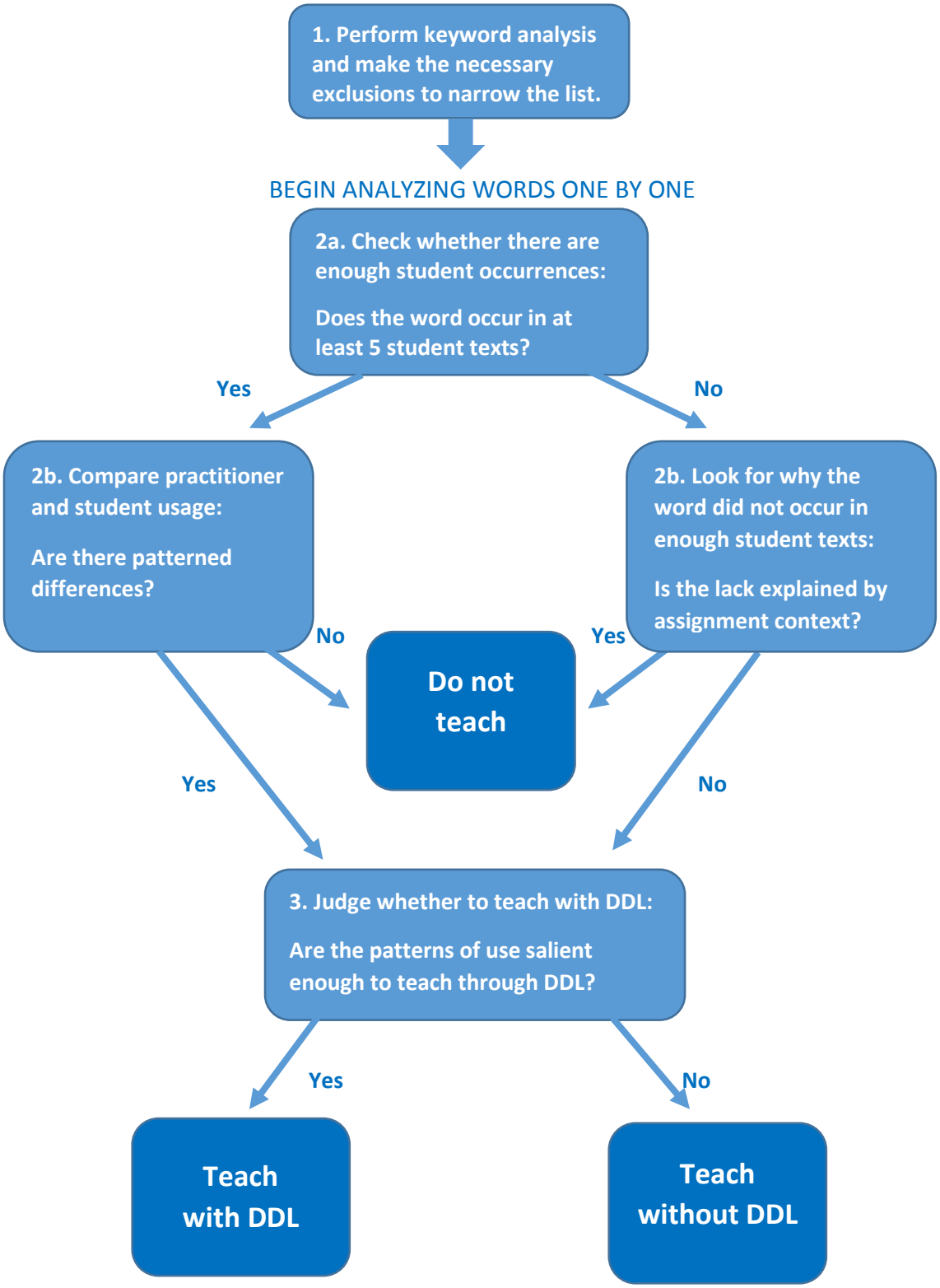


Figure 3.1 Flowchart summary of the systematic analysis process for determining specialized vocabulary to teach with DDL

files as the reference corpus. I created a COCA word frequency list ahead of time using AntConc and exported it as a text file.

Running keyword analysis in AntConc produced a list of keywords along with the rank (from highest to lowest keyness value), raw frequency, and keyness score for each word. I exported this keyword list to a plain text file and copied the list to an Excel spreadsheet for further analysis. I chose to cut off the keyword list at a minimum keyness score of 6.63 ($p < 0.01$) rather than a minimum keyness score of 3.84 ($p < 0.05$) in order to have a smaller number of keywords to work with and to limit the list to words that were more highly particular to civil engineering writing.

Because the keyword analysis generated a long list that included many words that were not useful for for my teaching goals, I needed a way to further narrow the list before analyzing the words in more depth. I limited my keyword list by coding and removing words according to seven categories (listed in Table 3.5 in the order that they were applied). The rationales for each exclusion category will be discussed in the following paragraphs.

The frequency and distribution requirements involved straightforward numeric cutoffs. I removed words with a raw frequency below 30 in order to ensure that the list represented words that are used often in civil engineering practice. I eliminated words that appeared in under 19 texts to help control for words that might only be frequent due to the specific context of a few projects (e.g., *eastbound*, *meters*, *vapor*). I eliminated words that appeared in under five of the seven registers to help control for words that were frequent due to particular specialties (e.g., *transportation*, *groundwater*, *masonry*).

Table 3.5 Summary and description of exclusion categories

Exclusion Category	Description
1. Frequency	Words that occurred fewer than 30 times in the corpus.
2. Non-words, acronyms, or anonymizations	Letters or groups of letters that were recognized as words by AntConc (e.g., *b*), acronyms (e.g., *DPE*), and anonymizations (e.g., *lastname*) that were added to the texts by the previous researchers to protect participants' identities.
3. Non-lexical words	Articles (e.g. *the*), prepositions (e.g. *for*), pronouns (e.g. *this*), be-verbs (e.g. *are*), modal verbs (e.g. *shall*), conjunctions (e.g. *and*), negation markers (e.g. *no*)
4. Text and register distribution	Words that were found in fewer than 19 texts (22% of the total texts) or fewer than 5 of the 7 registers.
5. Commonly used nouns	Basic nouns that are (1) likely to be familiar to most students even if English is not their first language and (2) do not change meaning significantly when used in engineering.
6. Proper nouns	Names of things, places, or organizations (e.g., *June, Oregon*).
7. Technical vocabulary	Words that are rarely used in general contexts and have clearly defined, specific, technical meanings (e.g., *silt, abutment, geotechnical*).

Words fitting the remaining categories were identified manually by reading the list and removing words that fit the categories' descriptions. Non-words and acronyms were excluded because they were not words. Anonymizations were excluded because they were not part of the original texts. I removed commonly used nouns because they were likely to be familiar to students and seemed less likely to have patterns of use that are particular to civil engineering. I removed technical vocabulary for several reasons. First, according to Mudraya (2006), technical words are often easier for students to acquire because they have a single meaning and do not have conflicting definitions or

usage patterns that the students may have learned in other contexts. Second, technical vocabulary is generally covered in course textbooks alongside new content; it is assumed that students do not know these words and that they will need to learn them in order to understand and describe specific concepts and information. Finally, technical vocabulary words are tied to specific engineering concepts that language instructors may not be able to understand or explain effectively.

I identified technical vocabulary by judging whether a word fit my definition of technical—having a clearly defined, specific, technical meaning and rarely being used in general, nontechnical contexts. For some words, this judgment was simple because I knew from experience that most people who are not engineers would be unfamiliar with the meaning of the words (e.g., *borings, geotechnical*). Other words were less clear-cut, but I determined that their use tends to be confined to specific contexts (e.g., *excavation, abutment, subsurface, silt*) or that they have a precise technical meaning (e.g., *diameter*).

3.3.2 Identifying words problematic for students. To identify words that were problematic for students, I looked at the words that were left from the practitioner-COCA keyword analysis after making exclusions and compared their usage between the practitioner and student corpora. I began with the word ranked highest by keyness score and worked down the keyword list, analyzing words one by one to decide whether they needed to be taught.

Because my goal was to assess the system and I did not have a required number of vocabulary words for a specific teaching application, I stopped after analyzing 45 words. The choice of 45 words was influenced by time constraints as well as the fact that this

number of words was sufficient for me to encounter a variety of word types (i.e., nouns, past- and present-participle verb forms, nontechnical and sub-technical words, etc.) and to accumulate enough words that did and did not need to be taught so that I could begin assessing the process. If any of these top 45 words was a noun with a singular or plural counterpart somewhere else in the list of keywords to be analyzed, I analyzed the singular and plural forms together as a single word.

To keep the comparative analysis structured and systematic, I considered five pieces of information for each of the 45 words to determine whether a word needed to be taught: comparative frequency, grammatical category, lexical bundles, collocations, and function. The following summarizes how I analyzed each of the five areas:

1. **Comparative frequency:** I began by checking which keywords were significantly more frequent in the practitioner corpus than in the student corpus in order to determine to what extent the students used the practitioner keywords. To establish significant frequency differences, I performed a keyword analysis of the practitioner corpus against the student corpus, using the same procedures as with the earlier practitioner versus COCA keyword analysis. I used the resulting keyword list to check which of the words I was analyzing had keyness values over 6.63 in the practitioner versus student keyword list. I marked each word with either "yes" or "no" in reference to whether they were used significantly less by students.
2. **Grammatical category:** I looked at concordance lines to determine and compare the grammatical category in the practitioner and student corpora. Specifically, I looked at word class for each word, and I made note of additional grammatical

features when relevant. For past and present participle verb forms, I noted whether the words were used as active- or passive-voice main verbs, as attributive adjectives, as noun postmodifiers, or in adverbial clauses. I made note of what grammatical categories a word was used in, whether the same grammatical categories were present for both groups, and whether there was a noticeable difference in which grammatical categories were dominant. If I found a noticeable difference (e.g., if practitioners usually used a word like *flow* as a noun and students usually used *flow* as a verb), it would indicate a need to teach the word.

3. **Lexical bundles**: Looking at lexical bundles allowed me to check whether formulaic phrases from the practitioner corpus were different from the word combinations that students were using. I examined lexical bundles using AntConc's cluster tool. I looked at bundles that were three to four words long with the keyword in any position in the bundle (e.g. first position: *observed in the*; middle position: *were observed in*). For practitioners, I set the minimum frequency (number of occurrences) at 10 and the minimum range (how many texts it occurred in) at five. For students, I set the minimum frequency and range at three, but because of the small size of the student corpus I also made note of any clusters that occurred at least twice.
4. **Collocations:** Looking at collocations allowed me to investigate words associated with the keyword I was studying. Seeing the group of top collocates from each corpus allowed me to look for trends to investigate (e.g., if most practitioner collocates were related to specific measurements and most student collocates were imprecise words like *hot* or *large*, this might indicate a difference in the level of

precision when using this keyword that could be investigated further). I examined top collocations using AntConc's collocate tool. The collocations were ordered by mutual information score, a statistical measure of collocational strength. I also set minimum frequency to eight for practitioners and two for students. I recorded the top collocates for each corpus and made note of any differences that were revealed after looking at the collocations in context.

5. **Function:** Function was determined through reading concordance lines and looking at the top lexical bundles and collocations in context. I interpreted function broadly to include the communicative purpose that the word served and to include the meaning that the word was intended to convey. An example of a functional difference can be seen with the word *required*; *required* was often used by practitioners to make recommendations while students often used *required* to state a requirement that they were aware of or needed to follow. Teaching the word *required* along with how to use it for this functional purpose would likely be valuable for preparing students to write for civil engineering practice.

I analyzed these five aspects of each word, kept notes on my findings in each area in a spreadsheet, and made a decision about teaching that weighed the five areas together. The process of deciding whether or not a word needed to be taught relied on making a judgment about whether or not there was a patterned difference between the usage. I weighed the five pieces of information for each keyword together to make this decision. I also considered the importance of the difference in deciding whether or not it constituted evidence that the word needed to be taught. For example, a difference in how

practitioners and students used a word functionally or the range of meanings that they conveyed with the word was generally considered enough evidence to mark a word as needing to be taught. In contrast, a difference in how many grammatical categories a word was used in was not typically considered enough evidence on its own and would require further evidence (such as a discernable pattern of differences in lexical bundles or collocations as well) to make a strong enough case for the word needing to be taught.

Since some of the keywords did not occur at all in the student corpus or occurred so few times that it was difficult to identify patterned differences, I decided not to comparatively analyze words that occurred in fewer than five student texts. Instead, I checked to see if there was a clear reason (such as the students' assignment context or subject matter) why the students did not use the word. If there was a clear reason why a student did not use a word, then I determined that there was not enough evidence to put it in the "teach" category and marked it as not needing to be taught. If there was not a clear reason why students did not use the word, then I marked it as needing to be taught and skipped to considering whether or not the word should be taught with DDL.

3.3.3 Identifying words well-suited to DDL. Following the decision of whether or not a word needed to be taught, I looked at the practitioner corpus to see whether or not the patterns of use were salient enough to be taught using DDL. I used pedagogical judgment to decide whether a data-driven approach would be the most efficient way to teach these words. I considered what needed to be taught (the gap between students and practitioners) and imagined looking at the corpus data through student eyes or imagined how I might design an activity to lead students to discover usage patterns.

My process with the word *samples* provides a clear example of how I determined whether a word was well suited for DDL. When comparing student and practitioner use of *samples*, I had noticed a difference in verb collocates and lexical bundles between the practitioners and students, particularly with the phrases *samples were collected* in the practitioner corpus and *samples were taken* in the student corpus. Once I decided that the word needed to be taught, I looked further into practitioner verbs that occurred in the phrase *samples were* _____. I listed the verbs that I saw, noted how many times they occurred, and grouped them into categories of things that civil engineers did with samples: obtaining samples, analyzing samples, and storing samples. Doing this allowed me to see how practitioners described their testing methodology, the details that they found important to report, and the verbs that they used to report these details. This investigational process seemed like something that students could replicate in a guided DDL activity, so I made notes for creating the DDL activity and marked the word *samples* as well suited to DDL (see Chapter 5 for the completed DDL activity with *samples*).

3.4 Evaluating the System

An important part of this study was assessing the effectiveness of the analytical system for determining vocabulary for DDL. I wanted to be able to describe the experience, reflect on whether the process was successful, and recommend changes for other ESP teachers who might be interested in using this process.

During the analysis process I took notes on what was helpful and not helpful, what discoveries I made, and anything that surprised me. I also recorded the reasons for

decisions that I made, including the reasons for keyword exclusions and for teaching decisions. I recorded the reason for each keyword exclusion that I made in order to give teachers an idea of how many words were eliminated for each criterion and which exclusion criteria were the most impactful. For teaching decisions, I marked the difference that was the main reason for the decision to teach each keyword (frequency, grammatical category, collocations, lexical bundles, or function) so that I could determine which factor was the most influential for my decisions.

I also performed several small tests to check whether elements in the analytical system were effective. To test whether the frequency requirement for keywords to occur at least 30 times in the practitioner corpus was too high, I examined the top 50 words that were excluded for frequency and counted how many would have been excluded for other reasons and how many would have ended up getting analyzed comparatively. To test whether significant differences in frequency between the practitioner and student corpora were influential on the teaching decision outcomes, I compared how many words in each teaching category—teach with DDL, teach without DDL, and do not teach—had significant practitioner-student frequency differences with how many did not have significant frequency differences. To test how time-consuming the comparative analysis process was, I timed the analysis of the last five of the 45 words to get an idea of how long it would take a teacher to judge words once familiar with the analysis and decision-making process.

Lastly, to help with this assessment and to demonstrate how the items in my list can be used, I created a DDL activity to teach the word *samples.* Creating this activity

helped me get a sense of whether or not the vocabulary words that I identified could be addressed in DDL activities. This activity can be found in Chapter 5.

3.5 Conclusion

This study involved using and assessing a corpus-based, analytical system for determining vocabulary to teach with DDL in an ESP context. I chose to focus on the field of civil engineering, and the Civil and Environmental Engineering program at PSU in particular, because of my experience with the Civil Engineering Writing Project. I performed my analysis by (1) comparing a corpus of expert civil engineering writing to a corpus of general English and creating a keyword list, (2) comparing the use of these keywords in the expert practitioner corpus and a corpus of low-proficiency student writing to find patterned differences, and (3) judging whether the words with patterned differences would be well suited for teaching through DDL. The next chapter presents the results of my corpus analysis and an assessment of the system for determining vocabulary.

Chapter 4: Results and Discussion

This chapter presents and discusses the findings of my study. The first part of the chapter presents the results of applying the analytical system to determine vocabulary to teach civil engineering students at Portland State University. The second part of the chapter discusses and assesses the effectiveness of the analytical system for determining specialized vocabulary for data-driven instruction.

4.1 The Vocabulary Results for Civil Engineering

This section includes the results of identifying words that were particular to civil engineering practice, identifying which words were problematic for students, and identifying which words were well suited to teaching through data-driven learning (DDL).

4.1.1 Identifying words particular to civil engineering. The keyword analysis comparing the practitioner corpus to the Corpus of Contemporary American English (COCA) produced a list of 2,641 words that were used significantly more in the civil engineering practitioner corpus ($p < 0.01$). The practitioner corpus contained only 7,497 word types, which means that about 35.2% of the word types were keywords. In comparison, the version of COCA that I used contained 945,695 word types—over 125 times as many word types as the practitioner corpus—meaning that the lexical diversity in the practitioner corpus was much smaller than in COCA. The small number of word types and the large percentage of keywords indicates that the vocabulary selection in the practitioner corpus was both narrow and specialized, which is consistent with the specific subject matter and purpose of the writing done in civil engineering practice.

I narrowed the list of 2,641 original keywords to a more manageable list of 201 words to consider for DDL teaching by excluding 2,440 words based on the criteria described in my methodology chapter. Table 4.1 displays the breakdown of these exclusions by number of words per exclusion category and the percentages of the total 2,641 keywords. The categories are displayed in the order that they were applied. Many of the keywords met the criteria for multiple exclusion categories, but I counted these words under the category where they first were excluded.

Table 4.1 Number of keywords excluded by exclusion criteria category

Exclusion Category	**Number of Words**	**Percentage of Total Keywords**
1. Raw frequency under 30	2,022	76.6%
2. Non-words, acronyms, or anonymizations	39	1.5%
3. Non-lexical words	31	1.2%
4. Failed distribution requirements	252	9.5%
5. Commonly used nouns	86	3.3%
6. Proper nouns	2	0.1%
7. Technical vocabulary	8	0.3%
Total Exclusions	**2,440**	**92.4%**
Remaining words after exclusions	**201**	**7.6%**

The words that failed to meet frequency and distribution requirements showed higher numbers than the other five categories. The following paragraphs discuss the kinds of words that were excluded based on frequency and distribution and an unexpected finding in the number of words excluded for being technical.

The words that failed frequency requirements included a large variety of words, many of which were technical (e.g., *dichloroethene, geophysical, stringer*) or uncommon because they were specific to certain projects or contexts (e.g., *Napa, snowmelt, ponds*).

This category also captured many acronyms (e.g., *PGE*), alphanumeric codes for samples or tests (e.g., *ha-16*), and some less common units of measurement (e.g., *deciliter, millimeter, microgram*).

The words in the distribution category were all excluded for being in fewer than 19 texts, although a number of them would have also failed for being in fewer than five registers. The words that failed distribution requirements also included many technical words (e.g., *basalt, liquefaction, subgrade*) and words that were context-specific (e.g., *wetland*, *Pease*, *Queens*). I also noted that many of the words that failed distribution were related to soil and geotechnical investigation (e.g., *boring, perchlorate, subgrade, clayey, silty*). The practitioner corpus was weighted towards geotechnical engineering, so it is not surprising that there were a large number of geotechnical vocabulary words present. The distribution requirements helped remove many of these geotechnical keywords.

I was initially surprised at the small number of words that were excluded as technical vocabulary. However, after seeing how many technical words were excluded for frequency or distribution failures, it made sense that only a few technical words had to be removed based on technicality alone. It seems that due to their specificity and narrow context of use, these words tended to be less frequent or well distributed than nontechnical words. Proper nouns as well, which were mainly related to specific projects, were almost all eliminated by frequency or distribution.

4.1.2 Identifying words problematic for students. After making the exclusions, I was left with a list of 201 words to be investigated by comparing practitioner and student use. As described in the methodology chapter, I analyzed only the top 45 words

to keep the research feasible while being able to observe a range of word types and outcomes.

After analyzing and comparing the practitioner and student use of the top 45 keywords with respect to frequency, grammatical category, collocations, lexical bundles, and function (as described in Chapter 3), I designated 18 words as needing to be taught, based on the differences between the two corpora (see the first two columns of Table 4.2). I designated 27 words as not needing to be taught based on a lack of evidence that the words were problematic to students. The words *surface* and *located* present typical cases where differences in usage, or a lack of differences, informed my decision to teach or not to teach. The following paragraphs illustrate how differences in meaning and context led to the decision to teach *surface* and how similar frequency, lexical patterns, meaning, and functions led to the decision not to teach *located*.

I identified the word *surface* as needing to be taught because of a patterned difference in the meanings and contexts of use between practitioners and students, a difference that also manifested in collocation, lexical-bundle, and grammatical-category differences. *Surface* was most often used by practitioners to indicate a position in space relative to the ground (i.e., surface-level as opposed to subsurface or higher than the ground). Students most often used the word *surface* to refer to the top or outside layer of an object. The following practitioner and student examples illustrate this contrast.

<u>Practitioners:</u>

a. In addition, *surface* drainage should be directed away from the top of slopes into swales or other controlled drainage devices.

b. This test was done at 10 feet below ground *surface*.

Table 4.2 Vocabulary list by decision to teach or not teach

Teach		Do Not Teach			
Rank	**Keyword**	**Rank**	**Keyword**	**Rank**	**Keyword**
7	existing	16	design	154	provided
14	proposed*	43	located	158	therefore
28/149	samples/sample	46/144	area/areas	164	west
29	observed	47/235	depth/depths	166	maximum
32	elevation*	49	report	168	assumed
36	approximately	54/78	pile/piles	173	east
40/165	slope/slopes	59	drainage*	174	shown
45	surface	82	retaining*	181	north
50	analysis	88/239	structure/structures	186	repair*
75	adjacent*	90	encountered*		
87	minimum*	91	constructed		
104	collected	95	temporary*		
117	required	102	additional		
133	performed	107	flow*		
142	using	110	fill*		
153	calculated	114/1018	results/result		
159	based	131	recommended*		
189	accordance*	148	lateral		
Total	**18 words**			**Total**	**27 words**

Note: Rank refers to the order of words by keyness score in the initial keyword analysis before exclusions were made.
*Indicates words with too low student frequency to be compared to practitioner use.

Students:

a. A water jet is shot at an impact *surface*, which deflects the water.

b. Slump cone was place on a dry (nonabsorbent *surface*), flat *surface*.

This contrast was also evident from differences in the lexical bundles and collocations from the two corpora. The most frequent practitioner bundles (e.g., *the ground surface*) and the top collocates (*ground, contamination, elevations*) for the word *surface* were mainly related to geotechnical investigation and discussion of ground features. The most

frequent student bundles (e.g., *surface of the*) and the top collocates (*nonabsorbent, flat, symmetrical, rough)* were related to tests that the students performed in their classroom labs, such as bending and applying pressure or blows to a sample of wood or metal and then recording changes in the surface appearance. Additionally, practitioners more frequently used *surface* as an attributive adjective (e.g., *surface conditions, surface drainage, surface elevation, surface soils, surface water*) than students did—which was likely due to the difference in meaning.

I determined that the word *located* did not need to be taught based on the fact that I did not find any patterned differences. *Located* was used with similar frequency in the practitioner and student corpus and with similar meaning and functions. The practitioners and students shared two of the same frequent lexical bundles—*located on the* and *located in the*—and the collocates did not reveal any striking differences. The following examples illustrate similar usage:

> Practitioner: The subject site *is located* in a light industrial area of San Bruno, California.
>
> Student: The station *is located* on the right side of a two lane, one way street.

During the comparative analysis, I came across 13 words that did not have enough student cases to be compared with the practitioner corpus (indicated with an asterisk in Table 4.2). My comparative analysis process involved looking for patterned differences between practitioner and student usage to use as evidence for whether or not a word needed to be taught, and it was not possible to establish patterns in so few student cases. For the words that were used in fewer than five student texts, I examined the student

corpus to see why they were not used and whether the lack of use was justified given the context of assignments. Five of these words with too few student cases were judged as needing to be taught, and eight were judged as not needing to be taught. The following examples with *temporary* and *adjacent* illustrate how I judged whether the lack of student use was explainable by context (i.e., not needing to be taught) or not explainable by context (i.e., needing to be taught).

The word *temporary* was only used in two student texts. The lack of student use appeared to be due to a lack of discussion of temporary structures rather than by students lacking the appropriate vocabulary. Because there was insufficient evidence indicating that it needed to be taught, temporary was put in the "do not teach" group. In contrast, the underuse of the word *adjacent*, which occurred in only two student texts, was problematic because there were many instances where it could have been used by students. Students used the alternate phrase *next to* in six student texts, while *next to* only occurred once in the practitioner corpus. The fact that students underused *adjacent* and often used *next to* instead of *adjacent* indicated to me that students were not familiar enough with the more register-appropriate term *adjacent* and that this term should be taught.

4.1.3 Identifying words well-suited to DDL. Of the 18 words needing to be taught, I identified 14 words as being well suited to teaching through DDL (shown in the first column of table 4.3) based on the saliency of the patterns in the practitioner corpus. For example, with the word *samples*, I had quickly noticed a difference between the top practitioner bundle (*samples were collected*) and collocate (*collected*) for *samples* and the

top student bundle (*samples were taken*) and collocate (*taken*) for *samples*. Investigating this difference further led me to notice a range of actions that practitioners performed, all indicated using passive voice: from gathering samples—*samples were collected* (27 times), *samples were obtained* (7 times), *samples were recovered* (3 times), *samples were taken* (1 time)—to analyzing samples—*samples were analyzed* (9 times), *samples were subjected to* (2 times), *samples were classified* (2 times)—to storing samples—*samples were sealed* (1 time), *samples were retained* (1 time). I found myself able to start sketching out a DDL activity to walk students through learning how practitioners describe the actions that they perform with samples (see Chapter 5 for the full DDL activity). Thus, I designated *samples* as well suited to teaching through DDL.

There were four words that displayed evidence of needing to be taught but that did not seem well suited to teaching through DDL. The usage patterns with the word *observed*, for example, were not easily noticeable by looking at concordance lines, collocations, or lexical bundles. I had marked *observed* as needing to be taught because of functional differences that I found, but the functional differences did not appear to be easily teachable through DDL. For more discussion of *observed*, see section 4.4.2. Ultimately, the analytical system's implementation with the civil engineering corpora was successful in yielding three lists (Table 4.3): words to be taught with DDL, based on patterned differences between practitioner and student usage and salient patterns that could be discovered by students; words to be taught without DDL, based on differences between practitioner and student usage but a lack of salient patterns; and words to not teach, based on a lack of differences between practitioner and student usage.

Table 4.3 Final teaching decisions for each of the 45 analyzed keywords

Teach with DDL		Teach without DDL		Do Not Teach	
Rank	**Keyword**	**Rank**	**Keyword**	**Rank**	**Keyword**
7	existing	29	observed	16	design
14	proposed*	32	elevation*	43	located
28/149	samples/sample	36	approximately	46/144	area/areas
40/165	slope/slopes	50	analysis	47/235	depth/depths
45	surface			49	report
75	adjacent*			54/78	pile/piles
87	minimum*			59	drainage*
104	collected			82	retaining*
117	required			88/239	structure/structures
133	performed			90	encountered*
142	using			91	constructed
153	calculated			95	temporary*
159	based			102	additional
189	accordance*			107	flow*
				110	fill*
				114/1018	results/result
				131	recommended*
				148	lateral
				154	provided
				158	therefore
				164	west
				166	maximum
				168	assumed
				173	east
				174	shown
				181	north
				186	repair*
Totals:	**14 words**		**4 words**		**27 words**

Note: Rank refers to the order of words by keyness score in the initial keyword analysis before exclusions were made.

*Indicates words with too low student frequency to identify patterned usage.

4.2 The Process of Determining Vocabulary for DDL—Was It Effective?

An important part of this study consisted of assessing a method for identifying vocabulary to teach with DDL. My goal was to find a process that English for Specific Purposes (ESP) teachers with moderate corpus experience could use for their own classes, comparing an expert specialized corpus and their own students' writing. My assessment and discussion of the analysis process will address the keyword analysis and keyword exclusion process, the comparative analysis process, and the judgment process.

4.2.1 The keyword and exclusion process. Overall, the keyword analysis and exclusion of extraneous keywords was effective. The keywords provided a crucial place to start when identifying vocabulary differences between the practitioner and student corpora. Starting with the keywords helped ensure that the vocabulary words were both important for civil engineering and particular to the discipline.

Keyword analysis was simple and quick to perform, and a visual scan of the keywords confirmed that many seemed strongly related to civil engineering. The raw keyword list output, however, was both too long (with an overwhelming 2,641 keywords) and was cluttered with words that were not useful for my teaching goals. High keyness value alone was not a good predictor of whether a word was worth spending time analyzing by comparing practitioner and student use: of the top 50 keywords by keyness score, for example, only 13 made it into the list of words to be analyzed comparatively.

The seven exclusion categories were effective in filtering the 2,641 keywords down to a more manageable 201 words. The exclusion process was more time-consuming than I had hoped, but most of this time was spent weeding out words manually for

categories like non-words, acronyms, and anonymizations; commonly used nouns; and non-lexical words. The frequency cutoff was the most efficient and powerful for cutting down the keyword list count, removing 2,022 of the 2,440 words that were excluded. The distribution cutoff was similarly straightforward and did not require judgment, but was more time-consuming than the frequency cutoff to apply since AntConc did not include an automatic way to filter by distribution.

Frequency and distribution were the most important exclusion categories for ensuring that the keywords were both commonly used and used in a range of contexts within the discipline. High keyness value was sometimes misleading as a measure of importance to civil engineering, as exemplified by *coflag*, which was excluded for low frequency, and *tank*, which was excluded for insufficient distribution. *Coflag* occurred only 16 times in the practitioner corpus; however, because it did not occur at all in COCA, the word had a high keyness value of over 247 (well above the 6.63 threshold for significance). *Tank* occurred 69 times in the practitioner corpus—which is about 47 times per 100,000 words compared to three times per 100,000 words in COCA, giving it a high keyness score of over 243; however, all of the cases of *tank* in the practitioner corpus were from the same eight texts, which indicated that the word was only important to the specific contexts of those texts and not to the general context of civil engineering. The exact parameters for frequency and distribution cutoffs that I used may not be directly transferrable to other ESP contexts because I chose the cutoffs based on the range of frequencies and the number of texts and registers in my study. Nevertheless, setting some kind of frequency and distribution limits is critical in order to remove keywords that have

keyness values above the significance threshold but are not of general usefulness to ESP students.

I also assessed whether the frequency cutoffs in my study were too high and whether they removed too many words with high keyness values that might have been useful vocabulary to teach. When I looked at the top 50 words by keyness value that had been excluded for their low frequency, I found that 40 out of these 50 excluded keywords would have been excluded for other reasons and that all of the words that were excluded for frequency had too low of a keyness value to have been included in the top 45 keywords that I ended up analyzing comparatively for this study. I also examined the words that would have been excluded if I had set the raw frequency cutoff higher at 40 or 50 occurrences (rather than 30, as I did in this study). I found that most of these words had ended up being excluded based on distribution or other criteria. Based on this finding, I concluded that the frequency cutoff was not too high, and that it may even have been efficient to set the frequency cutoff higher and to rely on the frequency and distribution filters alone to cut down the keyword list. This change would have saved time manually deciding whether to exclude commonly used nouns, technical words, etc. As long as the list of keywords was a manageable size, other undesirable keywords could have been dismissed quickly during the comparative analysis phase.

4.2.2 The comparative analysis process. I found that the process for comparing practitioner and student use was generally effective for finding usage differences to address in teaching. Considering five areas—comparative frequency, grammatical category, lexical bundles, collocations, and function—was effective for keeping my

analysis structured and contained. However, not all of these areas turned out to be equally useful to me for deciding whether or not a word needed to be taught. I found that differences in function or word meaning were the strongest determiners of whether a word needed to be taught; examining grammatical category, collocations, and lexical bundles was often useful for discovering differences in function or meaning; and comparative frequency did not impact my decisions—although, as I will discuss later, frequency differences may have been more important than I considered them at the time.

For each word that I marked as needing to be taught, I recorded the main reason for making that decision. The numbers of words for each decision (Table 4.4) showed that functional differences were the strongest decider (9 words), followed by lexical bundles and collocates (2 words), and that differences in grammatical category and frequency were not deciding factors in any of the teaching decisions that I made.

I also noticed during analysis that I tended to find functional differences more compelling than other differences. I found that functional differences alone were often sufficient for me to mark a word as needing to be taught, while differences in collocations, lexical bundles, grammatical category, or frequency were only strong enough in combination with each other. The case of the word *observed* illustrates this difference in how I weighted the factors.

The first difference that I saw with *observed* was that the practitioners used *observed* significantly more frequently than the students did. This information did not seem compelling on its own. The next difference that I saw was in grammatical category. Practitioners used *observed* mainly as a passive or active main verb, but they also used it as a noun postmodifier and as an attributive adjective. The student occurrences were all

Table 4.4 Number of vocabulary words for each reason to teach

Main reason for the decision to teach	**Words**
Differences in function or meaning	9
Differences in collocates and lexical bundles	2
Differences in grammatical category	0
Differences in frequency	0
Other*	2
Too few student cases and context did not explain the lack of use	5
Total words to teach	**18**

*Other included one case where the student uses indicated that they simply copied the word from the assignment template and a case with idiosyncratic student use that did not fall into any of the categories.

either passive or active main verbs. Examples of these category differences are shown in Table 4.5. This lack of grammatical variety in the student corpus, while different from the practitioner writing, was not strong enough justification that the word needed to be taught—especially considering that there were only eight student occurrences compared to the practitioners' 245 occurrences. Given the small number of student occurrences, it was not problematic that the students only used *observed* in the two grammatical categories that were most common in the practitioner corpus.

Function was a more significant factor in the decision to teach *observed*. I noticed from scanning concordance lines and looking at bundles and collocates in context that practitioners often used *observed* to report their findings from site visits before making recommendations, as in the following example:

> Major cracks are *observed* at several locations especially at the rear-west portion of the building. The house was not found structurally stable and therefore it is recommended to demolish the house.

Table 4.5 Examples of observed by grammatical category in practitioner and student writing

Practitioner Writing	Student Writing
Passive main verb: Black hydrocarbon staining was *observed* at 2 ft bgs in the UST excavation.	Passive main verb: Also, the data collected was not the same lab that was *observed*, which means there could be unknown variables.
Active main verb: We *observed* silty sandy gravel (GP) with occasional cobble.	Active main verb: This is where I *observed* the difference in the railing for the Portland Streetcar and the Max.
Noun postmodifier: The basalt boulders *observed* on the western shoreline 150 feet downstream of the proposed crossing likely indicate that the bedrock is near the surface in this area.	
Attributive adjective: The *observed* subsurface and surface conditions are consistent with a landslide extending from the ridge ±700 feet south of the bridge site, north-northwest to near US Highway 30.	

Some of the student examples also used *observed* to report details from site visits, as in the following example:

> When we went the cully plaza building we *observed* that the building has some destruction on the site and some of it is clean.

However, this example sentence is part of a description of the project background which provides information about the site rather than the findings of an investigation that will be used as the basis for a recommendation. Other student occurrences made more casual observations (example a) or even described the procedure for a lab report (example b):

a. This is where I *observed* the difference in the railing for the Portland Streetcar and the Max. … Another interesting thing that I *observed* was the concrete that was used to create these roadways for the Max and Streetcar.

b. It will be *observed* as well in order to study the fracture area.

I found this difference in functions compelling evidence that *observed* should be taught.

Although functional differences, when discovered, tended to hold more weight on their own than differences in the other areas, looking at grammatical categories, collocations, and bundles was useful for observing how the practitioners and students used the words. Collocations and lexical bundles tended to provide the clearest leads for discovering functional and other differences, so I will discuss these two areas in more detail.

Collocates turned out to be useful in many cases for identifying themes to investigate further. With the word *encountered*, for example, the first seven collocates were all related to taking soil samples through boring: *formation, sand, soils, borings, groundwater, fill, below*. When I looked at these collocates in context, I confirmed that the word *encountered* was primarily used to report findings from geotechnical investigation. The word *using* exemplifies how differences in practitioner and student collocations revealed underlying functional differences. Looking at the collocates of *using* from both groups, *determined* and *performed* from the practitioner corpus stood out as being unlike any of the student collocates. On further investigation, I found that practitioners often used *using* to document how they performed a test or determined certain results, thus establishing credibility, as in the following examples:

a. The stability analysis was *performed using* XSTABLE slope stability analysis software.
b. The 50-year storm event water surface elevation was *determined using* HEC-RAS.

In contrast, the students often used *using* to demonstrate knowledge to their instructor, describe how something functioned, or write up a standard lab procedure that they followed, as in the following student examples:

a. A LVDT is used to measure small movements or deformations of specimens. … This system detects small voltages when a deformation occurs in the specimen. *Using* a voltmeter an output can be read of the specimen. *Using* a derived relationship with voltage and length a deformation can be found indirectly.
b. This data can be obtain by *using* excels maximum function to locate, however due to the fluctuations of the data, it may not be entirely accurate.

Lexical bundles turned out to be similarly useful for highlighting functional differences and, at times, other differences. For example, the practitioner and student bundles for *performed* were noticeably different:

Practitioners: *should be performed* (occurred 11 times)

Students: *performed in this, performed to determine, test performed in, to be performed* (each phrase occurred only twice)

The presence of *should* in the practitioner bundle—and the lack of it in the student bundles—led me to find that the practitioners often used *performed* when making

recommendations, while students more commonly used perform to describe actions they had completed, as in the following examples:

> Practitioner: Excavation to expose the native subgrades *should be performed* using a straight-edged bucket without traversing the subgrade.
>
> Student: The lab *performed* was a Tension Test, *performed* on two different coupons, #1 being light and aluminum; #2 being dark and steel.

Practitioners also used *performed* to describe actions they completed, but I found that, even in these cases, there were differences in grammatical choices: practitioners were more likely than the students to use active voice when describing actions they had completed, as in the following examples:

> Practitioner: *We performed* laboratory tests on selected soil samples to check the corrosion potential to subsurface metal structures.
>
> Student: The lab *was performed* using two different coupons.

In the case of *performed*, examining lexical bundles led me to discover differences in both functional and grammatical behavior.

I had hoped that grammatical categories, collocations, and bundles would reveal more simple, teachable differences (e.g., if practitioners almost always used a particular word in the passive voice and students mostly used it in the active voice, it might be worth teaching this lexico-grammatical association). I had also hoped that collocations and lexical bundles would reveal simple word associations and formulaic sequences that I could teach. However, the fact that investigating grammatical category, collocations, and lexical bundles revealed deeper functional differences rather than simple, surface-level differences was perhaps more valuable, as it helped me connect the language civil

engineers use to the work that civil engineers do. Discovering associations between vocabulary words and disciplinary functions—especially with words like *encountered, using,* and *performed* that are not explicitly related to civil engineering concepts—is incredibly valuable for ESP teachers.

The last area of practitioner-student comparison to discuss is frequency. During my analysis, frequency did not factor highly into my teaching decisions, and I initially concluded that it should be eliminated from the system for determining ESP vocabulary to teach. When I compared the numbers of words that were used significantly less by students in each of the teaching decision categories, it was clear that significant frequency differences were not a useful predictor of my teaching decision (Table 4.6).

Table 4.6 Number of keywords analyzed by significant frequency differences and teaching decisions

	Teach with DDL	**Teach without DDL**	**Don't Teach**	**Totals**
Students used significantly less	10	4	20	**34**
Students did not use significantly less	4	0	7	**11**
Totals	**14**	**4**	**27**	**45**

On further reflection, however, I found that differences in frequency may have been more meaningful than I originally thought. It is worth investigating cases where students used a word much less than practitioners—such as the word *observed*, which students used only eight times (which is about two times per 10,000 words, compared to practitioner use of about 17 times per 10,000 words). The lack of use of *observed* may have indicated student avoidance of the word—perhaps because of unfamiliarity—or may

have indicated a lack of understanding about the importance of reporting observations in civil engineering.

I briefly explored the question of why students did not use *observed* as frequently as practitioners, and I saw that students used two alternate words that were less common in the practitioner corpus: *noticed* (student frequency: about 1.5; practitioner frequency: about 0.2, normed per 10,000 words) and *saw* (student frequency: about 1 per 10,000 words; not used by practitioners). Neither of these words, however, approached the frequency of practitioner use of *observed*, which suggests that students may have an underdeveloped understanding of the importance of observation in civil engineering in addition to underusing the word *observed*. This word presents an excellent opportunity to teach the word, usage, and importance of the concept for civil engineering together.

It is often easier for teachers to address the problems they see than to address what students avoid using (see, e.g., Schachter, 1974). It would be worth looking more closely at frequency differences in the future and giving these differences more weight in the decision of whether a word needs to be taught.

4.2.3 The judgment process—whether to teach and whether to use DDL. Making judgments about whether or not a word needed to be taught and whether a word was well suited for DDL was more difficult and more time-consuming than I had hoped. I had also hoped that the process of analyzing five pieces of information would make decisions more clear-cut—that is, if no patterned differences between practitioner and student use were evident after following this process, then the word could be dismissed. The process, while helpful, did not make decisions entirely clear-cut. There was still a

subjective and sometimes difficult decision about whether to teach and whether to use DDL that had to be made after investigating each word.

I found, however, that the analysis and judgment process sped up considerably as I became more familiar with the system and had a clearer idea of what kind of usage differences I would consider to be evidence for a word needing to be taught and how salient the usage patterns needed to be to consider the word well-suited to DDL. I timed the comparative analysis and judgment portion for last five of the 45 keywords that I analyzed so that I could give teachers an idea of how long this process would take once they were familiar with it. Table 4.7 shows these five words, the time I spent examining practitioner and student use, the decisions I made, and the reasoning behind the decisions.

These five words were likely faster than some other words because the first four did not require a decision about whether or not they were well suited for DDL. Additionally, *accordance* did not require a full comparative analysis because there were too few student cases; instead I only had to investigate the student corpus to see if there were contexts where the accordance should have been used and then investigate the salience of patterns in the practitioner corpus to decide whether or not the patterns were salient enough for DDL. *Accordance* had straightforward use patterns that were quick to find since this word tends to be used in a very formulaic phrase: *in accordance with the*. Although these five words were likely faster than average and were analyzed when I was most familiar with the process, the times that I recorded can help give teachers a sense of how much time this process can take.

Table 4.7 The last five keywords analyzed with their analysis times and teaching decisions

Keyword	Minutes	Decision	Reasons
east	16	Do not teach	I did not find any patterned differences except that practitioners used it significantly more than students because of different context demands.
shown	11	Do not teach	I did not find any patterned differences.
north	10	Do not teach	I did not find any patterned differences except that practitioners used it significantly more than students because of different context demands.
repair	2	Do not teach	There were too few student occurrences to analyze and the lack of student use could be explained by context.
accordance	10	Teach with DDL	There were too few student occurrences to analyze and the lack of student use could not be explained by context.
Total	**47 min**		

As a whole, the process of compiling two specialized corpora, performing keyword analysis, excluding keywords based on criteria, comparing practitioner and student use, and making decisions about whether to teach and whether the words were well suited to DDL was fairly involved and time-consuming. This method would likely not be worth performing for a just a lesson or two. However, this method could be a valuable tool for planning vocabulary to systematically include in DDL activities in ESP classes. The ability to identify specialized vocabulary words that are important for an ESP context and customized to the current needs of students would be valuable and worth the time for course-level or program-level planning.

4.3 Conclusion

Testing this analytical system for the context of civil engineering was a productive experience for learning more about civil engineering vocabulary, confirming that the system was feasible, and making observations that would be useful for future implementation. These were my most notable observations:

1. The overall process of determining specialized vocabulary for teaching through DDL was involved and time-consuming but would be worth performing when planning an ESP course or program, in order to systematically include vocabulary instruction.
2. Ensuring that the keywords met frequency and distribution requirements was an important step in this process because high keyness value alone did not always indicate usefulness in a variety of civil engineering contexts.
3. Discovering functional differences associated with specific vocabulary words was a valuable contribution of the comparative analysis.
4. More investigation and attention should be given to student underuse of vocabulary in future applications of this system.

The next chapter concludes this thesis by discussing how these findings can be applied in language instruction and suggesting directions for future research.

Chapter 5: Conclusion

This chapter concludes my thesis with a summary of my findings, applications for the context of civil engineering education, broader applications for context-specific language instruction, some limitations of the study, and directions for future research.

5.1 Summary of Findings

The goal of this study was to investigate the usefulness of an analytical system for determining English for Specific Purposes (ESP) vocabulary to teach with data-driven learning (DDL). Testing out this analytical system with a limited number of words (45), I identified 14 vocabulary words that were important for writing in a civil engineering context, were problematic for students, and were well suited to teaching through DDL.

In my assessment, I concluded that the system was successful overall in meeting my goals of (1) identifying a list of words that were significantly more frequent in civil engineering writing than in general writing, (2) identifying words that showed patterned differences between practitioner and student use, and (3) identifying words where the usage patterns in the practitioner corpus were salient enough to teach through DDL. The system was complicated and time-consuming to implement but would be worthwhile for compiling a set of vocabulary words that are likely to be most useful to ESP students.

I also made several useful observations. The first observation was that setting limits for minimum frequency and distribution in the practitioner corpus was important. Setting these limits, in addition to requiring that the words be significantly more frequent in civil engineering writing than in general writing, helped ensure that the words were common in civil engineering writing and were used in a range of contexts.

Another important observation was that looking for differences in grammatical category, collocation, and lexical bundles often revealed more compelling functional differences between the practitioner and student writing. These functional differences were not what I was originally looking for, but it ended up being highly valuable for linking engineering functions to specific vocabulary words that students needed to learn.

A third significant finding was that during the student-practitioner comparison, I tended to overlook differences in frequency that may have been important indicators of words needing to be taught. After looking more closely at the word *observed*, which had a large difference in practitioner and student frequency, it became clear that student underuse of *observed* was an important difference that should have been given more weight.

5.2 Applications for Civil Engineering Writing Instruction

The immediate application of this study was identifying a set of vocabulary words that would be useful for creating DDL activities for civil engineering students with low-proficiency writing skills at Portland State University. Of the words that I looked at for this exploratory study, I identified these 14 words as a starting point for creating DDL activities:

existing, slope/slopes, minimum, performed, based, proposed, surface, collected, using, accordance, sample/samples, adjacent, required, calculated

In addition to identifying these words, my corpus analysis also revealed usage patterns and functions associated with these words, which will be relevant for the creation of DDL

activities. To illustrate how these words could be taught with DDL, I created an example activity for teaching the word *samples*.

The teaching activity for *samples* is a hands-on DDL activity designed for students to perform using AntConc (Figure 5.1). Students load the civil engineering practitioner corpus that I used for this study into AntConc, perform the searches described in the activity, and write answers to the questions indicated. This activity could be used during classroom time or assigned as homework if students have enough experience with similar activities. This activity fits into the category of guided induction described by Smart (2014) and discussed in Chapter 2. The activity is structured so that students will find the exact results that I planned for them to find. I designed this activity with the assumption that students have been trained in using AntConc. Further direction would be necessary for students who are not trained in corpus use.

While this activity is built around *samples*, it also focuses on other vocabulary words that collocate with *samples* and explores the meaning created when the words are combined. One of those collocates is *collected*, which was also one of the vocabulary words that I found to be well suited for teaching with DDL. *Collected* also frequently occurred as a postmodifier of *samples*, as in the following example:

> All soil *samples collected* from the borings (primary and duplicate samples) were analyzed for perchlorate (EPA Method 314) by the EPA Region 9 Laboratory in Richmond, California.

This noun-postmodifier relationship could be explored in a second DDL activity. Vocabulary words with close relationships, such as *samples* and *collected*, can be taught together effectively. Teaching them together can also help to highlight important

Vocabulary Word: Sample(s)

Test your knowledge:

1. This activity focuses on how to talk about samples in civil engineering. You are probably already familiar with the word ***samples***. To test your knowledge, think of a specific context where you use samples and write a sentence using the word ***sample or samples***.

Investigate with the corpus:

2. To see cases of *sample* and *samples* together, type *sample|samples* into the search bar. Read through some of the concordance lines to see examples of how civil engineers use this word. Write down at least one example that is interesting to you.

3. In the collocate tab, search for words that come after *sample|samples*.
   window span from 0 to 3R
   min. collocate frequency: 10

   What are the top four words? Write them here.

4. Click on each of the first four words to see them in context. What do you notice about most of them? (Hint: are they in active voice?) Write down one example that seems common.

5. In the concordance tab, search for the phrase *samples were*. You should get 62 hits.

6. Sort the results and make a list of all the *–ed* or *–en* verbs that come after *samples were*. Write down each phrase below and how many times it occurs. (The first one is done for you.)

Phrase with -ed verb	**How many times?**	**What kind of action? (for question #7)**
samples were collected	27	Collecting
(samples were analyzed)	*(9)*	*(Analyzing)*
(samples were obtained)	*(7)*	*(Collecting)*
(samples were classified)	*(25)*	*(Analyzing)*
(samples were recovered)	*(2)*	*(Collecting)*
(samples were retained)	*(3)*	*(Storing/moving)*
(samples were sealed)	*(1)*	*(Storing/moving)*
(samples were subjected to)	*(1)*	*(Analyzing)*
(samples were taken)	*(2)*	*(Collecting)*
(samples were transported)	*(1)*	*(Storing/moving)*

7. For each phrase in the chart from question #6, decide which category of action it fits into. (Look at the phrases in context if you aren't sure.)
   a. **Collecting** the samples
   b. **Analyzing** the samples
   c. **Storing/moving** the samples

Figure 5.1 Sample teaching activity for *samples* (continued on next page)

8. What does “samples were classified” mean?

9. Which *–ed* verbs were the most common for each action? Circle them in the chart.

10. Were any of these *–ed* verbs new or surprising for you? Write down one sentence from the corpus that was unusual or different from how you usually use the word *samples*.

Apply what you learned:

11. Look back at the sentence that you wrote in question #1. Is there anything you would change now that you have observed how civil engineers typically use the word *samples*? Revise your original sentence or write a new sentence applying what you found in the civil engineering corpus.

Figure 5.1 Sample teaching activity for *samples* (continued from previous page)

functions in civil engineering, such as the function of providing critical information about how and where samples were collected.

As discussed in Chapter 2, the need for disciplinary literacy instruction goes beyond second-language learners. The group of students that my vocabulary list was intended for included both first- and second-language English speakers who have a low proficiency in writing. Portland State University does not currently offer civil engineering ESP classes or civil engineering disciplinary writing classes, and so there is great opportunity to offer support to civil engineering students with both first- and second-language English backgrounds. The words that I identified as needing to be taught could be developed into a set of vocabulary materials to be used in workshops, tutoring, or other supplemental writing instruction.

5.3 Broader Applications for Teaching Disciplinary Literacy

The analytical system that I developed for identifying specialized vocabulary makes use of specialized corpora to tailor vocabulary choices to a specific language context and a specific group of learners. This makes it an excellent tool for teachers to

apply in their own particular contexts. I envisioned that teachers, particularly ESP instructors, would create a corpus of their own students' papers and either select a previously-compiled expert corpus or compile an expert corpus for their discipline. This system can help ESP or other writing instructors who are teaching language skills for a field or discipline that they do not have first-hand experience in.

The findings from applying this system can also benefit content instructors, such as engineering faculty. Content teachers are often not consciously aware of what words are important to their disciplines unless the words' meanings are explicitly tied to the content area. My analysis of *encountered, using, and performed*—words that are used in many contexts and easy to overlook—revealed engineering functions tied to these seemingly basic words. Findings like these can be shared with content teachers to raise their awareness of key vocabulary that their students need to know. Content teachers are best equipped to teach students about the actions and functions performed in their discipline—and with increased awareness of vocabulary, they can point out words during instruction that students should use when discussing certain concepts or reporting particular information.

My findings from testing this analytical system in the context of civil engineering had a number of implications for teachers who are interested in using this system for their classes.

1. This analytical system was time-consuming. It is important to plan accordingly and to keep the comparative analysis between the practitioner and student corpora moving quickly. Using a timer to monitor the time spent on each step and each

keyword, as I did with the last five keywords, may be helpful for avoiding getting bogged down in any one step or keyword.

2. Frequency and distribution limits on keywords were important for making sure that the vocabulary words were regularly useful in a range of civil engineering contexts. I would strongly recommend including some kind of frequency and distribution limits in future implementation. Additionally, while AntConc did not allow me to automatically set distribution limits, other concordancing software does have this capability and would be worth trying. Other studies have also used key-keywords—words that are keywords in each register when registers are individually compared with a general corpus—as a way of regulating distribution. Other exclusion filters beyond frequency and distribution may or may not be useful, depending on individual teaching goals.
3. Looking at grammatical category, collocations, and lexical bundles was most productive as a way to find functional differences. I would recommend looking at these three areas with an eye out for functional differences. Examining these areas in the same order for each word also adds structure to the comparative analysis.
4. Differences in frequency could have been investigated in more depth and given more weight than I did. I would recommend that teachers look at student underuse more thoroughly since it may be an indication of avoidance or other problems that keep students from using words as frequently as experts typically do.

5.4 Limitations and Future Research

This study looked at only one discipline when assessing the analytical system for determining vocabulary for teaching with DDL, and my results cannot be generalized past civil engineering. An important area for future research is to apply this system in other disciplines and ESP contexts to see if it is effective in other contexts. My results and findings about the effectiveness of the system were also based on the judgments and experience of a single teacher. It would be beneficial to study whether another teacher replicating my methodology with the same data would make similar teaching decisions and conclusions. In future studies, including judgments from multiple teachers and checking interrater reliability would strengthen this system.

Another limitation of this study had to do with the method of identifying students with low-proficiency writing skills. Using grammar error counts as a measure of writing proficiency did not guarantee that my student corpus accurately represented the students who would benefit most from vocabulary instruction or that my results could be generalized to the wider population of Portland State University civil engineering students who need writing support. My results also cannot be generalized to engineering students beyond Portland State University. Replication of this study using corpora from other civil engineering contexts would be useful in the future.

Additionally, future research could benefit from involving content experts, such as engineering faculty or practitioners, in the process to help with decisions and interpreting results. Content experts would be best qualified to identify technical vocabulary that are typically covered in textbooks, to confirm whether words identified

by the system are in fact important to their discipline, and to identify disciplinary functions being performed in writing samples.

Future research could also explore the impact of teaching the words from this study on students' vocabulary and writing development. It would be useful to see how effective instruction of the words determined in this study would be for civil engineering students at PSU. It would also be useful to replicate the present study and teach the vocabulary words that are identified, in order to present a fuller evaluation of the analytical system for selecting vocabulary.

Another area for expansion would be to adapt this system or develop a similar system for determining other language features to teach in disciplinary contexts with DDL. A number of studies have looked at teaching grammar features (e.g., Vannestål & Lindquist, 2007; Smart, 2014) and even genre features (e.g., Weber, 2001) through DDL, so it would be useful to consider ways for teachers to systematically determine what features should be taught and which would be well suited to teaching through DDL.

5.5 Final Thoughts

One of the greatest challenges for a teacher is knowing what to teach—what will make the greatest impact on students, what will make the most efficient use of class time, and what will prepare students for the challenges they face outside of the classroom. ESP teachers are tasked with an important and difficult job: refining language instruction so that class time addresses features of language use that are most relevant to the discourse communities students plan to participate in. Especially for new teachers or teachers with

limited experience with the disciplinary contexts that they are addressing, having tools to guide teaching decisions is critical.

Corpus research has contributed a number of excellent teaching and reference tools for ESP teachers, such as DDL, but there are still many more problems and questions to address. My hope for this analytical system to determine vocabulary for DDL is that it adds one more tool to the ESP teacher's toolbox and that it empowers teachers as they work to prepare students to competently participate in new disciplinary contexts.

References

Anthony, L. (2015). AntConc (Version 3.3.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net

Anthony, L. (2016). AntConc (Windows, Macintosh OS X, and Linux) Build 3.4.4 [Help file]. Retrieved from http://www.laurenceanthony.net/software/antconc/releases/AntConc344/help.pdf

Barlow, M. (2002). MonoConc 2.2 [Computer Software]. Houston, TX: Athelstan.

Biber, D., & Conrad, S. (2009). *Register, genre, and style.* Cambridge University Press. doi: 10.1017/CBO9780511814358

Boulton, A. (2009). Testing the limits of data-driven learning: Language proficiency and training. *ReCALL, 21*, 37–54. doi: 10.1017/S0958344009000068

Boulton, A. (2012). Corpus consultation for ESP: A review of empirical research. In A. Boulton, S. Carter-Thomas & E. Rowley-Jolivet (Eds.), *Corpus-informed research and learning in ESP: Issues and applications* (pp. 261–291). Amsterdam: John Benjamins. doi: 10.1075/scl.52.11bou

Brandl, K. (2008). *Communicative language teaching in action: Putting principles to work.* Upper Saddle River, NJ: Pearson Prentice Hall.

Brozo, W. G., Moorman, G., Meyer, C., & Stewart, T. (2013). Content area reading and disciplinary literacy: A case for the radical center. *Journal of Adolescent & Adult Literacy, 56*, 353–357. doi: 10.1002/jaal.153

Centre for English Corpus Linguistics. (n.d.). The Louvain corpus of native English essays. Retrieved from http://www.learnercorpusassociation.org/resources/tools/locness-corpus

Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal corpora. *English for Specific Purposes, 35*, 30–40. doi: 10.1016/j.esp.2013.11.004

Chujo, K. & Utiyama, M. (2006). Selecting level-specific specialized vocabulary using statistical measures. *System, 34*, 255–269. doi: 10.1016/j.system.2005.12.003

Civil Engineering Degree. (2017). Career options for civil engineers. Retrieved from http://www.civilengineeringdegree.org/career-options-for-civil-engineers

Conrad, S. (2000). Will corpus linguistics revolutionize grammar teaching in the 21st century? *TESOL Quarterly, 34*, 548–560. doi: 10.2307/3587743

Conrad, S. (2014). Expanding multi-dimensional analysis with qualitative research techniques. In T. B. Sardinha & M. V. Pinto (Eds.), *Multi-dimensional analysis 25 years on: A tribute to Douglas Biber* (pp. 275–98). Amsterdam: Johns Benjamins. doi: 10.1075/scl.60

Conrad, S. (2015). Register variation. In D. Biber & R. Reppen (Eds.), *Cambridge handbook of English corpus linguistics* (pp. 309–329). Cambridge: Cambridge University Press. doi: 10.1017/cbo9781139764377

Conrad, S. (in press). A comparison of practitioner and student writing in civil engineering. *Journal of Engineering Education, 106*(2). doi: 10.1002/jee.20161

Conrad, S. (to appear). The use of passives and impersonal style in civil engineering writing. *Journal of Business and Technical Communication*.

Conrad, S., & Pfeiffer, T. (2011). Preliminary analysis of student and workplace writing in civil engineering. Proceedings of the 2012 American Society for Engineering Education Conference. Retrieved from http://www.asee.org/search/proceedings

Conrad, S., Pfeiffer, T., & Szymoniak, T. (2012). Preparing students for writing in civil engineering practice. Proceedings of the 2012 American Society for Engineering Education Conference. Retrieved from http://www.asee.org/search/proceedings

Coxhead, A. (2013). Vocabulary and ESP. In P. Paltridge & S. Starfield (Eds.), *The handbook of English for Specific Purposes* (pp. 115–132). Oxford: John Wiley & Sons. doi: 10.1002/9781118339855.ch6

Davies, M. (2008–). The corpus of contemporary American English: 450 million words, 1990–Present. Available from http://corpus.byu.edu/coca/

Donesch-Jezo, E. (2010). Corpus concordancing in teaching academic discourse writing to medical students. Proceedings of the International Conference ICT for Language Learning. Retrieved from http://conference.pixel-online.net/ICT4LL2010/conferenceproceedings.php

Ellis, R. (2002). Methodological options in grammar teaching materials. In E. Hinkel & S. Fotos (Eds.), *New perspectives on grammar teaching in second language classrooms* (pp. 155–179). Mahway, NJ: Lawrence Erlbaum Associates. doi: 10.4324/9781410605030

Ellis, R., Basturkmen, H., & Loewen, S. (2002). Doing focus-on-form. *System, 30*, 419–432. doi: 10.1016/S0346-251X(02)00047-7

Flowerdew, L. (2003). A combined corpus and systemic-functional analysis of the problem-solution pattern in a student and professional corpus of technical writing. *TESOL Quarterly*, *37*, 489–511. doi: 10.2307/3588401

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System*, *32*, 301–319. doi: 10.1016/j.system.2004.04.001

Gilquin, G., Granger, S., & Paquot, M. (2007). Learner corpora: The missing link in EAP pedagogy. *Journal of English for Academic Purposes, 6,* 319–335. doi: 10.1016/j.jeap.2007.09.007

Gollin, J. (1998). Deductive vs. inductive language learning. *ELT Journal, 52,* 88–89. doi: 10.1093/elt/52.1.88

Graham, D., & Osment, C. (2013). Interactive web-based learning of corpus-generated phrases. *AsiaCALL Online Journal, 9,* A1–A13. Retrieved from http://asiacall.info/acoj/acoj-2014/

Harklau, L. (2000). From the "good kids" to the "worst": Representations of English language learners across educational settings. *TESOL Quarterly, 34,* 35–67. doi: 10.2307/3588096

Hartig, A. J., & Lu, X. (2014). Plain English and legal writing: Comparing expert and novice writers. *English for Specific Purposes, 33,* 87–96. doi: 10.1016/j.esp.2013.09.001

Henderson, R., & Hirst, E. (2006, November). How sufficient is academic literacy? Re-examining a short-course for "disadvantaged" tertiary students. Paper presented at the 2006 AARE Conference, Adelaide, Australia. Retrieved from https://eprints.usq.edu.au/1598/2/Henderson_Hirst_AARE_2006_PV.pdf

Hou, H. I. (2014). Teaching specialized vocabulary by integrating a corpus-based approach: Implications for ESP course design at the university level. *English Language Teaching, 7,* 26–37. doi: 10.5539/elt.v7n5p26

Hu, M., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*, 403–430. Retrieved from https://eric.ed.gov/?id=EJ626518

Huang, Z. (2014). The effects of paper-based DDL on the acquisition of lexico-grammatical patterns in L2 writing. *ReCALL, 26*, 163–183. doi: 10.1017/s0958344014000020

Johns, T. F. (1991). Should you be persuaded – Two samples of data-driven learning materials. *ELR Journal*, *4*, 1–16. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.477.809

Kaewpet, C. (2009). Communication needs of Thai civil engineering students. *English for Specific Purposes*, *28*, 266–278. doi: 10.1016/j.esp.2009.05.002

Kennedy, C. & Miceli, T. (2001). An evaluation of intermediate students' approaches to corpus investigation. *Language Learning & Technology 5*(3), 77–90. Retrieved from http://llt.msu.edu/vol5num3/pdf/kennedy.pdf

Liu, D., & Jiang, P. (2009). Using a corpus-based lexicogrammatical approach to grammar instruction in EFL and ESL contexts. *The Modern Language Journal*, *93*, 61–78. doi: 10.1111/j.1540-4781.2009.00828.x

Lee, D., & Swales, J. (2006). A corpus-based EAP course for NNS doctoral students: Moving from available specialized corpora to self-compiled corpora. *English for Specific Purposes, 25*, 56–75. doi: 10.1016/j.esp.2005.02.010

Marinov, S. (2013). Training ESP students in corpus use - Challenges of using corpus based exercises with students of non-philological studies. *Teaching English with*

Technology, *13*, 49–76. Retrieved from http://cejsh.icm.edu.pl/cejsh/element/bwmeta1.element.desklight-768dfe18-527c-41fd-a4ed-491d8d828070

Marzá, N. E. (2014). A practical corpus-based approach to teaching English for tourism. *International Journal of Applied Linguistics and English Literature*, *3*, 129–136. doi: 10.7575/aiac.ijalel.v.3n.1p.129

Moje, E. B. (2008). Foregrounding the disciplines in secondary literacy teaching and learning: A call for change. *Journal of Adolescent & Adult Literacy*, *52*, 96–107. doi: 10.1598/jaal.52.2.1

Mudraya, O. V. (2004). Need for data-driven instruction of engineering English. *IEEE Transactions on Professional Communication*, *47*, 65–70. doi: 10.1109/tpc.2004.824296

Mudraya, O. V. (2006). Engineering English: A lexical frequency instructional model. *English for Specific Purposes*, *25*, 235–256. doi: 10.1016/j.esp.2005.05.002

Nation, P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review, 63*, 59–82. doi: 10.3138/cmlr.63.1.59

Nelson, S. (2000). Teaching collaborative writing and peer review techniques to engineering and technology undergraduates. *30th ASEE/IEEE Frontiers in Education Conference*, *2*, S2B 1–5. Retrieved from http://fie-conference.org/fie2000/papers/1422.pdf

Önder, N. (2014). Using corpus data to teach collocations in medical English. *Journal of Second Language Teaching & Research*, *3*, 37–52. Retrieved from http://pops.uclan.ac.uk/index.php/jsltr/article/view/237

Portland State University. (2017). Maseeh college of engineering & computer science: Civil & environmental engineering. Retrieved from https://www.pdx.edu/cee

Salager-Meyer, F. (1994). Hedges and textual communicative function in medical English written discourse. *English for Specific Purposes*, *13*, 149–170. doi: 10.1016/0889-4906(94)90013-2

Schachter, J. (1974). An error in error analysis. *Language learning*, *24*, 205–214. doi: 10.1111/j.1467-1770.1974.tb00502.x

Schmidt, R. W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*, 129–158. doi: 10.1093/applin/11.2.129

Shanahan, T., & Shanahan, C. (2012). What is disciplinary literacy and why does it matter? *Topics in Language Disorders*, *32*, 7–18. doi: 10.1097/TLD.0b013e318244557a

Smart, J. (2014). The role of guided induction in paper-based data-driven learning. *ReCALL*, *26*, 184–201. doi: 10.1017/s0958344014000081

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. *Language Learning Journal*, *36*, 139–152. doi: 10.1080/09571730802389975

Sun, Y. C., & Wang, L. Y. (2003). Concordancers in the EFL classroom: Cognitive approaches and collocation difficulty. *Computer Assisted Language Learning*, *16*, 83–94. doi: 10.1076/call.16.1.83.15528

Swarts, J., & Odell, L. (2001). Rethinking the evaluation of writing in engineering courses. *31st ASEE/IEEE Frontiers in Education Conference*, *1*, T3A 25–30. Retrieved from http://icee.usm.edu/ICEE/conferences/FIEC2001/papers/1283.pdf

Vannestål, M. E., & Lindquist, H. (2007). Learning English grammar with a corpus: Experimenting with concordancing in a university grammar course. *ReCALL*, *19*, 329–350. doi: 10.1017/s0958344007000638

Weber, J. J. (2001). A concordance-and genre-informed approach to ESP essay writing. *ELT Journal*, *55*, 14–20. doi: 10.1093/elt/55.1.14

Wingate, U. (2015). *Academic literacy: The case for inclusive practice.* Bristol: Multilingual Matters.

Winsor, D. A. (1990). Engineering writing/writing engineering. *College Composition and Communication*, *41*, 58–70. doi: 10.2307/357883

Woodward-Kron, R. (2008). More than just jargon–the nature and role of specialist language in learning disciplinary knowledge. *Journal of English for Academic Purposes*, *7*, 234–249. doi: 10.1016/j.jeap.2008.10.004

Yalvac, B., Smith, H. D., Troy, J. B., & Hirsch, P. (2007). Promoting advanced writing skills in an upper-level engineering class. *Journal of Engineering Education*, *96*, 117–128. doi: 10.1002/j.2168-9830.2007.tb00922.x

Yoon, H. (2008). More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning & Technology, 12*(2), 31–48. Retrieved from http://llt.msu.edu/vol12num2/yoon.pdf

Yunus, K., & Awab, S. (2012). The effects of the use of module-based concordance materials and data-driven learning (DDL) approach in enhancing the knowledge of collocations of prepositions among Malaysian undergraduate law students.

International Journal of Learning, 18, 165–181. Retrieved from http://ijl.cgpublisher.com/product/pub.30/prod.3335

Yunus, K., & Awab, S. (2014). The impact of data-driven learning instruction on Malaysian law undergraduates' colligational competence. *Kajian Malaysia, 32,* 79–109. Retrieved from http://web.usm.my/km/vol32_supp1_2014.html